On Computational Stylistics: Mining Literary Texts for the Extraction of Characterizing Stylistic Patterns

Mohamed Amine Boukhaled

Université Pierre et Marie Curie

École Doctorale Informatique, Télécommunications et Électronique

Laboratoire d'Informatique de Paris 6

On Computational Stylistics: Mining Literary Texts for the Extraction of Characterizing Stylistic Patterns

Présentée par

Boukhaled Mohamed Amine

Thèse de Doctorat en Informatique

Dirigée par le Professeur **Jean-Gabriel Ganascia**

Présentée et soutenue publiquement le *13 septembre 2016*

Devant un jury composé de:

BEAUDOUIN, VALÉRIE	Directrice d'Études	Examinatrice
GANASCIA, JEAN-GABRIEL	Professeur	Directeur de Thèse
MARSALA, CHRISTOPHE	Professeur	Examinateur
MINEL, JEAN-LUC	Professeur	Rapporteur
POIBEAU, THIERRY	Directeur de Recherche	Rapporteur
SOLDANO, HENRY	Maître de Conférences	Examinateur

Abstract

The present thesis locates itself in the interdisciplinary field of *computational stylistics*, namely the application of statistical and computational methods to the study of literary style. Historically, most of the work done in computational stylistics has been focused on lexical aspects especially in the early decades of the discipline. However, in this thesis, we tackle a different linguistic level than lexicon. Our focus is put on the *syntactic aspect of style* which is quite much harder to capture and to analyze given its abstract nature.

As main contribution, we work on an approach to the computational stylistic study of classic French literary texts based on a *hermeneutic* point of view, in which discovering interesting linguistic patterns is done without any prior knowledge. More concretely, we focus on the development and the extraction of complex yet computationally feasible stylistic features that are linguistically motivated, namely *morpho-syntactic patterns*.

Following the hermeneutic line of thought, we propose a *knowledge discovery process* for the stylistic characterization with an emphasis on the syntactic dimension of style by extracting relevant patterns from a given text. This knowledge discovery process consists of two main steps, a *sequential pattern mining* step followed by the application of some *interestingness measures*. In particular, the extraction of all possible syntactic patterns of a given length is proposed as a particularly useful way to extract interesting features in an exploratory scenario. Clearly the proliferation of patterns and the difficulty for humans to make sense of huge amount of results are major obstacles to this approach. Therefore, we use interestingness measures in this scenario to treat and reduce such large quantities of patterns in order to identify the most relevant ones. We propose, carry out an experimental evaluation and report results on three proposed interestingness measures, each of which is based on a different theoretical linguistic and statistical backgrounds.

The analyzed results of the experimental evaluation indicate that the presented techniques are fairly effective in extracting interesting syntactic patterns, especially if we take into account the unsupervised nature of this process. This seems particularly promising as a *computer-assisted literary analysis tool* to support linguists and literary researchers in their stylistic analysis.

In the case study, we apply the proposed methodology to a more specific application than the general stylistics framework: the study of *theatrical stylistic characterization* by analyzing the voice of Molière's characters in terms of distinguishing syntactic patterns.

Finally, we present a computational stylistic tool partly which concretizes our research work called *EReMoS*. The goal of *EReMoS* is to provide linguists and literature researchers with a computer-assisted stylistic tool conceived as a web application capable of extracting and manipulating syntactic patterns through a simple, fast and ergonomic user interface.

Keywords: computational stylistics, sequential data mining, knowledge discovery, text mining, morpho-syntactic pattern, interestingness measure, correspondence analysis, outlier detection, theatrical stylistic characterization, computational authorship study, EReMoS.

Résumé

Titre de la thèse en Français:

De la Stylistique Computationnelle: Fouille de Textes Littéraires pour l'Extraction de Motifs Stylistiques Caractérisants

La présente thèse se situe dans le domaine interdisciplinaire de la *stylistique computationnelle*, à savoir l'application des méthodes statistiques et computationnelles à l'étude du style littéraire. Historiquement, la plupart des travaux effectués en stylistique computationnelle se sont concentrés sur les aspects lexicaux. Cependant, dans notre thèse, nous abordons un niveau linguistique différent du lexique. En effet, l'accent est mis sur l'*aspect syntaxique du style* qui est beaucoup plus difficile à capturer et à analyser étant donné sa nature abstraite.

Comme contribution principale, dans cette thèse, nous travaillons sur une approche à l'étude stylistique computationnelle de textes classiques de littérature française d'un point de vue *herméneutique*, où découvrir des traits linguistiques intéressants se fait sans aucune connaissance préalable. Plus concrètement, nous nous concentrons sur le développement et l'extraction des *motifs morphosyntaxiques*.

Suivant la ligne de pensée herméneutique, nous proposons un processus de découverte de connaissances pour la caractérisation stylistique accentué sur la dimension syntaxique du style et permettant d'extraire des motifs pertinents à partir d'un texte donné. Le processus de découverte de connaissances proposé consiste en deux étapes principales, une étape *d'extraction de motifs séquentiels* suivi de l'application de certaines *mesures d'intérêt*. En particulier, l'extraction de tous les motifs syntaxiques possibles d'une longueur donnée est proposée comme un moyen particulièrement utile pour extraire des caractéristiques intéressantes dans un scénario exploratoire. Il est clair que la prolifération de ces motifs et les difficultés que rencontre l'homme pour donner du sens à d'énormes quantités de résultats constituent des obstacles majeurs à cette approche. Par conséquent, dans notre cas, nous utilisons des mesures d'intérêt pour traiter et réduire ces grandes quantités de motifs afin d'en identifier les plus pertinents. Nous proposons, évaluons et présentons des résultats sur les trois mesures d'intérêt proposées, basée chacune sur un raisonnement théorique linguistique et statistique différent.

Les résultats de l'évaluation expérimentale indiquent que les techniques présentées sont assez efficaces pour extraire des motifs syntaxiques intéressants, en particulier si l'on tient compte de la nature non supervisée de ce processus. Par conséquent, ces techniques semblent particulièrement prometteuses comme *outil d'analyse littéraire assistée par ordinateur*.

Dans l'étude de cas, nous appliquons la méthodologie proposée à un domaine plus spécifique que le cadre général de la stylistique, soit l'étude de *caractérisation stylistique théâtrale* en analysant le discours des personnages de Molière en termes de motifs syntaxiques distinctifs.

Enfin, nous présentons un outil de stylistique computationnelle appelé *EReMoS* concrétisant en partie notre travail de recherche. Le but d'*EReMoS* est de fournir aux linguistes et aux chercheurs en littérature un outil de stylistique assisté par ordinateur et conçu comme une application web capable d'extraire et de manipuler des motifs syntaxiques grâce à une interface utilisateur simple, rapide et ergonomique.

Mots clés: Stylistique computationnelle, fouille de données séquentielles, découverte de connaissances, fouille de textes, motif morphosyntaxique, mesure d'intérêt, analyse des correspondances, détection des cas aberrants, caractérisation théâtrale, étude computationnelle de paternité, EReMoS.

Aknowledgement

First of all, I would like to thank my advisor Prof. Jean-Gabriel Ganascia for giving me the opportunity to pursuit this thesis, for his precious advices, and for his full support and encouragement. I am very grateful for the mutual respect climate that he established during all these years.

Of course, I would like to express my deepest gratitude to the honorable jury members of my defense who accepted to evaluate my work; to Christophe Marsala, Henry Soldano, Jean-Luc Minel, Thierry Poibeau and Valérie Beaudoin.

I also express my dearest gratitude towards all past and current members of the ACASA team, as well as all the members of Labex OBVIL.

This PhD thesis has been a wonderful experience in such an attractive scientific field. I feel grateful to all the persons who have in one way or another guided me along this journey.

I am very thankful to my long date friends too, those in Paris, those in Algiers, and those now spreading all around the world.

Finally, I want to dedicate the last words of acknowledgment to my parents and my beloved sisters. Thank you so much for your love and support.

To my family.

Publications

The work performed during the thesis was at the origin of several publications and scientific contributions. The list below presents a summary of these contributions:

Boukhaled, M.A., & Ganascia, J.-G. (2014). Probabilistic Anomaly Detection Method for Authorship Verification. In S. I. Publishing (Ed.), *2nd International Conference on Statistical Language and Speech Processing, SLSP 2014* (Vol. 8791, pp. 211–219). Grenoble, France: Springer.

Boukhaled, M.A., & Ganascia, J.-G. (2014). Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules. In *The 11th International Workshop on Natural Language Processing and Cognitive Science* (pp. 115–122). Venice, Italy: DE GRUYTER.

Boukhaled, M.A. (2015). Une méthode non supervisée pour la vérification d'auteur à base d'un modèle gaussien multivarié. In *10es Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI)* (pp. 525–533). Paris, France: ARIA.

Boukhaled, M.A., Frontini, F., & Ganascia, J.-G. (2015). A Peculiarity-based Exploration of Syntactical Patterns: a Computational Study of Stylistics. In *Workshop on Interactions between Data Mining and Natural Language Processing DMNLP'15 ECML/PKDD 2015 Workshop* (pp. 31–40). Porto, Portugal.

Boukhaled, M.A., Frontini, F., & Ganascia, J.-G. (2015). Une mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques. In *22ème Conférence sur le Traitement Automatique des Langues Naturelles*. Caen, France.

Boukhaled, M.A., Sellami, Z., & Ganascia, J.-G. (2015). Phoebus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires. In *22ème Conférence sur le Traitement Automatique des Langues Naturelles*. Caen, France.

Frontini, F., **Boukhaled, M.A.**, & Ganascia, J.-G. (2015). Moliere's Raisonneurs: a quantitative study of distinctive linguistic patterns. Corpus Linguistics 2015, Lancaster, UK.

Oudni, A., **Boukhaled, M.A.**, & Bourgne, G. (2015). Analyse des relations et des dynamiques de corpus de textes littéraires par extraction de motifs graduels. In *24ème Conférence sur la Logique Floue et ses Applications, LFA2015*. Poitiers, France.

Riguet, M., Jolivet, V., & **Boukhaled, M.A.** (2015). « Cohérence sémantique »: l'apport des algorithmes de représentation vectorielle des mots. In *8es Journées Internationales de Linguistique de Corpus*. Orleans, France.

Boukhaled, M.A., Frontini, F., Bourgne, G., & Ganascia, J.-G. (2015). Computational Study of Stylistics: a Clustering-based Interestingness Measure for Extracting Relevant Syntactic Patterns. *International Journal of Computational Linguistics and Applications. 06*(01), 45–62.

Contents

Chapter 1. General Introduction

1.1. Context and Motivation

Digital technologies are completely changing many aspects of our daily lives. Human beings find themselves more and more relying on digital technologies to accomplish their tasks. For example, digital technologies have fundamentally changed the printing industry. Technologies and devices such as eBook had big impacts on the cultural aspect of people's life by altering the way they read texts or consume information. This impact has spread across many others fields besides industry such as economics and science as well.

Actually, digital technologies, enabling fast and robust computing, transformed the science field by changing the way scientists use to engage in their research activities. Indeed, it is becoming evident that research is increasingly dependent on digital technology. The impact of this new development on science varies depending on disciplines and research areas. For instance, nature sciences are nowadays fundamentally and unquestionably reliant on digital technologies and computation unlike other fields such as the humanities.

Digital humanities, known in its early days as "computing in the humanities", or "humanities computing" is precisely the research field that covers a range of methods and approaches at the intersection of both computing and disciplines of the humanities (Siemens & Schreibman 2013).

The lexical shift that the field of digital humanities has known actually reflects the important development that changed how the computer and the computation are perceived in this discipline through its history. Schnapp et al. (2009) summarize this development in their Digital Humanities Manifesto 2.0 by saying: "*The first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is qualitative, interpretive, experiential, emotive, generative in character*".

Moreover, digital humanities, which began as a term of consensus among a relatively small group of researchers, is now one of the fastest and promising growing research areas (Kirschenbaum 2012). The importance of using and developing computational approaches in the humanities is not an isolated phenomenon but rather a part of a much bigger shift known as the computational turn (Berry 2011). In addition to the humanities in general and literary analysis in particular, the computational turn is increasingly reflected across a number of disciplines, including but not limited to the arts and social sciences (Lazer et al. 2009).

In this context, our thesis work comes as part of the effort started in the Labex OBVIL (the observatory of literary life). This laboratory intends to develop and exploit resources offered by computer applications to examine French literature. It promotes scientific research in the field of digital humanities by bringing together researcher from both literary and social sciences on the one hand, and computer scientist and engineers on the other hand.

Text, as a support for literary productions and as opposed to other data formats such as images and videos, has traditionally by far been the most manipulated data type by computers. Actually, even some of the earliest computers had the capabilities to process texts, which resulted in a long tradition of developing computational text analysis tools for research field such as linguistics and literary analysis.

In fact, the use of computational approaches in the study of literary texts has a long-standing tradition. If we consider the word computational in its etymological sense of counting, we can date back such approaches prior to the era of computers (Lutoslawski 1898, Mosteller & Wallace 1963). Again in the field of linguistics (and stylistics as well), the application of quantitative methods to the analysis of style and genre dates back to the beginnings and continues today (Leech and Short 2007, Semino and Short 2004, Biber 2011, Mahlberg 2013). It is nevertheless undeniable that in recent years quantitative methods have moved out of the margins and into the forefront of literary studies, thanks to the availability of large quantities of digitized texts and to the success that data mining methods have had in the identification of historical trends in literature (Moretti 2005, Jockers 2013) that would have been hard to spot to a naked eye. While the advantages of computational methods are evident when treating huge corpora, they exist also for smaller ones; single books or even parts of them, as we shall see in this dissertation, may disclose interesting and new insights when analyzed from a different and new perspective.

1.2. Thesis Topic and Objectives

Historically, most of the work done in stylometry is focused on lexical aspects especially in the first decades of the discipline. Moreover, the few works that deal with the syntactic aspect of style were either rule-based or more focused on syntactic characteristics that can be analyzed without the need for any advanced natural language processing tools.

Indeed, the present work locates itself in the long tradition of stylometry, namely the application of statistical methods to the study of literary style (Holmes 1998). However in our thesis, we tackle a different linguistic level than lexicon. Our focus is put on the syntactic aspects of style which are much harder to capture and to analyze given their abstract nature. In fact, stylometric methods have often been applied to tackle issues of authorship attribution, but more recently a different discipline has evolved out of this field, one on which our work is centered, namely *computational stylistic* in which computational method are applied as an analytic tool for the investigation of significant stylistic traits characterizing a literary work, an author, a genre, a period, etc.

Research in computational stylistics is typically associated with the study of the authorial signal, namely the identification of a given author's typical traits through a comparison of his or her work to that of others (known as individual style as opposed to functional style). Previous studies have often privileged the analysis of discrete units, typically of words. In this thesis, we propose and describe a computational stylistics methodology that combines the bottom up extraction of morpho-syntactic patterns, with a type of statistical assessment methods called interesting-ness measures, and we apply this methodology to the study of stylistic characterization, namely to automatically finding characterizing morpho-syntactic traits in some author's writings. Generally speaking, our main working hypothesis in this thesis is that more complex linguistic features are used in a more conscious and controlled way, and thus, when some of them are strongly over-used

or under-used in an author's novel with respect to other ones or exhibit a peculiar behavior and distribution, this may be taken as a possible interesting stylistic trait.

As main contribution, in our thesis we have worked on an approach to the computational stylistic study applied on classic French literature texts based on a hermeneutic point of view where discovering interesting linguistic patterns is done without any prior knowledge or explicit a priori classification.

More concretely, we focus on the development and the extraction of complex yet computationally feasible stylistic features that are linguistically motivated, namely *morpho-syntactic patterns*.
Based on the literature review that we have conducted as an important and fruitful part of our thesis, we claim that computational stylistic methods need to be grounded in the hermeneutic unsupervised paradigm rather than on the classification-based one. Following this line of thought, we propose a knowledge discovery process for stylistic characterization with an emphasis on the syntactic dimension of style by extracting relevant patterns from a given text. The proposed knowledge discovery process consists of two main steps, a sequential data mining step followed by the application of some interestingness measures. We propose, evaluate and report results on three interestingness measures, each of which is based on a different theoretical linguistic background.

Our aim is to conceive and develop a framework that is meant to assist linguists and literary researchers in studying the syntactic style, and in extracting meaningful linguistic patterns from the text they are interested in. More concretely, it is meant to support the stylistic textual analysis, especially from a syntactic perspective, by:

1) Verifying the degree of importance of each extracted linguistic pattern

2) Automatically inducing a list of linguistic features that are significant and representative for an author's work

3) Allowing to read the text in a controlled and systematic manner by providing the ability to read the results, sort and filter them, and view them within the context of the text as well

1.3. Thesis Organization

The remaining parts of this thesis are organized as follows:

Chapter 2 provides the general framework within which our work fits, namely computational stylistics. It gives the reader an overview of this discipline including the approaches and the methodologies used to carry out computational stylistic activities, along with an overview of some other related fields. It also reports on woks that we consider being very important and influential to the contribution that we have made during the thesis.

Chapter 3 represents the core part of our thesis contribution. In that chapter, we present the approach considered for the extraction of relevant stylistic patterns and we give details about the proposed knowledge discovery process. We present the proposed interestingness measures used to assess the relevancy of those extracted patterns as well.

In **Chapter 4**, we report on the experimental evaluation and discussion of the resulting patterns.

Chapter 5 is meant to be a case study of the proposed knowledge discovery process, in which we focus on the stylistic analysis of memorable protagonist characters of prose plays written by Molière. Our aim in that chapter is to study the stylistic singularity that Molière gives to his protagonists in its syntactic form.

Chapter 6 concludes this thesis by presenting a summary of contributions and results, highlighting the limits, and providing some directions for future work.

Finally, in **Appendix A** our software contribution (EReMoS web application) is presented, while **Appendixes B** and **C** present our contributions in the field of computational authorship attribution and verification.

Chapter 2. Literature Review on Computational Stylistics

Every one of us has a unique personality and manners of communicating messages. These manners translate into what we call the stylistic behavior (Allport 1961). Basically, the stylistic behavior covers just about all of what we do and what we say: the way we talk, the way we walk, how we express emotions, how we use gestures or how we dress for instance. All these ways constitute the personality style that differentiates each one of us from others, and reflects the different dimensions of someone's personality. One of the very important phenomena that can clearly illustrate the stylistic behavior is language. Actually, people are very attentive to the way a message is transmitted as much as to its content.

In a very basic point of view, it can be seen as how people choose their words then put them all together to create a message. And in this case as well, as there is much about the linguistic content of the message, there is also a linguistic style chosen to convey it. Following this idea, text, which constitutes an important unit of language and which can be seen as the material form of language (Kress 1988), is made up of two characteristics that are its propositional content and its decorative and communicative form.

It is important not only to acquire the propositional content of a text but also its communicative form and how these propositions are advanced by the author. In fact, fully understanding texts require not only comprehending the content in its abstract form, but also going beyond that by taking a look into more other deep and relevant aspects such as the semantic content, the communicative and decorative effects, and the interaction between them covered by the style of the text.

Stylistics and stylometry are traditionally two research fields interested in studying the notion of style in texts. In the most general understanding of these fields, stylistics is basically the sub discipline of linguistics interested in the study, the analysis and the interpretation of texts on a stylistic basis. On the other hand, stylometry is that discipline which, for the stylistic study of texts, integrates and relies upon statistical procedures to achieve its goal (Grzybek 2014).

In the 1940s, the modern computers appear. Stylometry especially has gained shortly after that a considerable development thanks to these new devices that allowed computerizing the statistical analysis and performing data analysis of textual data at large scale. This new development was responsible for the emergence of computational stylistics.

In this first chapter, we present a literature review of the computational stylistic field. What one should keep in mind is that given its interdisciplinary nature, computational stylistics is in practice a very scattered discipline. Thus, it is actually worth mentioning that this chapter is not meant to be an exhaustive listing of the works done in this discipline so far, but rather an introductory chapter trying to bridge the gap between the different components of computational stylistics by making abstraction of the technical details.

At first, in Section 2.1 of this chapter we smoothly start by introducing the reader to the notion of style, which is in fact a central point to our thesis, by focusing on a technical perspective of it. In Section 2.2 we go into the details of computational stylistics, its history, definition, goals and challenges. After that, we discuss the shared ground of computational stylistics by introducing some strongly related fields such as corpus stylistics for instance (Section 2.3), and then we take a bigger perspective in Section 2.4 by talking about the different approaches that have emerged in this field. Section 2.5 is focused on one of the successful and worth discussing applications of computational stylistics, namely computational authorship attribution. Since our thesis is centered on the analysis of the syntactic dimension of style, Section 2.6 gives an overview of the computational analysis of the syntactic aspect of style. Finally, in Section 2.7 we report in details

on three strongly related works that we consider being very important and influential to the work carried out during the thesis.

2.1. The Notion of Style

When writing a text, besides deciding the propositional content, the author has to decide on many other necessary aspects related to his writing that go along with the advanced propositional content. As a language producer, the author must take many linguistic decisions in order for the text to be correctly formed. These decisions vary on different linguistics levels. Since language is a very regulated and complex phenomenon, we can arguably assume that these decisions are not randomly taken, but rather chosen in a specific and defined manner that embeds additional information in the text. This additional information and specificity lead the text to exhibit a particular style of writing that may be recognizable by people who are familiar with it. They also reflect the author's intent to convey the effect related to his particular style. Together, propositional content and stylistic effects end up characterizing not only a single piece of written text but also a manner of writing in its general form.

Among many other factors, those elements make the style play an essential role in the content and in the meaning addressed by the author to the reader. In this section, we do not intend to discuss all the possible definitions and interpretations of style, which has been in fact extremely differently defined in the literature. Yet, it seems worthwhile to give at least two definitions of the writing style precisely. Well, as we have said so far, style may be roughly defined as the manner in which something is expressed, as opposed to the content of a message (Argamon et al. 2005). But more deeply, one can notice in the literature two complementary and overlapping points of view for defining the notion of style from a technical (processional) perspective: as a set of choices, and as a set of variations.

For the choice-based point of view, we assume that language is a sequence of options and choices and that every propositional content or idea can be expressed in many different ways (Halliday 1978). Some authors tend to prefer a certain language options to others. Moreover, they tend to be recursive when making those preferences. Such preferred selections from the whole set of choice possibility allowed by the language environment differentiate one author from another.

The possibilities offered by a language environment vary on different linguistics levels. For example, the first and the most basic one corresponds to the lexical level both in its qualitative and quantitative part, that is to say which words are chosen by an author and how many times each. Then come more deep and complex levels such as the syntactic level for instance in which words can be linked into sentences based on the syntactic possibilities allowed by the syntactic rules of the language[1].

Sanders (1977) summarized and formalized this idea in his *principle of choice* which is precisely the basis of his stylistic theory. He claimed that style is the result of choices made by an author from a range of possibilities offered by the language system.

[1] We will come back to this idea in Subsection 2.3.2 of this chapter when we talk about natural language processing and computational linguistics.

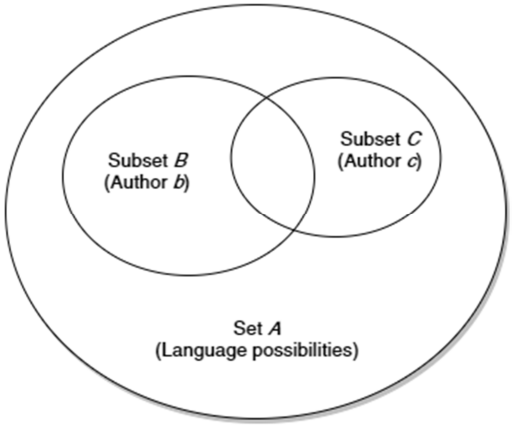

Figure 1. Sanders' principle of choice

The *principle of choice*'s idea can be illustrated in Figure 1. It clearly shows how this principle differentiates the writing of a given author from another one. In this picture, the set A represents all the options and possibilities offered by a language environment, while the sets B and C represent the fragment of options chosen by two different authors to express an idea, author b and author c respectively. Of course such representation can be generalized to include as many authors as we want.

The second point of view that we can take when addressing the notion of style is the variation-based one. In this line of thinking, DiMarco & Hirst (1993) pointed out that: "*style is created through subtle variation, seemingly minor modulations of exactly what is said, the words used to say it, and the syntactic constructions employed, but the resulting effect on communication can be striking*".

Stylistic variation, that can be considered as a stylistic indicator, depends on several factors such as author profile (age, gender, education, etc.) and competence, genre, communicative context, expected characteristics of the intended audience and so on (Argamon et al. 2005).

What these two points of view presented above about the notion of style have in common is that they focus more on a technical linguistic perspective. Riffaterre (1971) in his book *Essais de stylistique structural* gives for instance a definition that emphasizes more the rhetorical perspective of style: *"Le style est compris comme un soulignement ajouté à l'information transmise, sans altération de sens. [...]. Le style est la mise en relief qui impose certains éléments de la séquence verbale à l'attention du lecteur, de telle manière que celui-ci ne peut les omettre sans mutiler le texte et ne peut les déchiffrer sans les trouver significatifs et caractéristiques (ce qu'il rationalise en y trouvant une forme d'art, une personnalité, une intention, etc.). Ce qui revient à dire que le langage exprime et que le style met en valeur*"[2].

[2] "The style is seen as an emphasis added to the information transmitted without alteration of meaning. [...]. The style is the highlighting that imposes some elements of the verbal sequence for the attention of the reader, so that he cannot omit them without mutilating the text and cannot decipher them without finding them significant and characteristics (what he rationalizes by finding in it an art form, a personality, intention, etc.). This is to say that language expresses and style highlights" [translation provided by the thesis' author]

2.2. From Style to Computational Stylistics

2.2.1. Brief History of the Interaction between Stylistics and Statistics

Studying style is a very ancient quest. Already in the Antiquity, scholars from the ancient Greek have been studying it as part of the rhetorical study of texts. The idea was that the rhetorical form achieved through a careful choice of words and syntactic structures is responsible for reflecting the writer's thoughts and intentions (Corbett 1973). Starting from the fifth century B.C, Greek rhetoricians such as Corax of Syracuse developed methods for the systematic instruction of written texts. Later, scholar such as Isocrates and Aristotle established very influential standards of rhetoric (DiMarco & Hirst 1993).

The interaction between mathematics and statistics on the one hand and language study including style study on the other hand is very ancient tradition as well[3]. The Alexandrian grammarians had already listed the hapax legomena of Homer, and Masoretes had counted every word of the Bible (Guiraud 1960). In the 8th century The Arab linguistic scholar Al-Farahidi developed what is known as one of the first binary encoding system to study and recognize phonetic patterns in the classic Arabic poetry, which led him to the identification of the 16 Arabic poetry meters known and respected till our days (Khalaf et al. 2011). The work done by Italian humanist Lorenzo Valla in the 15th century on proofing that the *Donation of Constantine* was a forgery is considered to be an important historical work on stylometry (Grzybek 2014).

This tradition of interactions between statistics and mathematics on the one hand and language study (including style-related studies) or texts as a material form of language on the other hand, known commonly as textometry[4], continued to exist in the following centuries, but it didn't surpass the stage of counting some linguistic units (phonetic and lexical units in most cases). Of course, this statistical information about those linguistics units was valuable in the sense that it was used afterward to build some reflections about the studied texts, however it was lacking both in term of generality (the counting was performed in one single text) and foundations (it wasn't founded on some established linguistic theory).

[3] The interaction between mathematics and language study did not go only in one way (mathematics → language study), it was actually a bidirectional interaction. Maybe the most illustrative example to cite in this matter is the Markov chains. In fact, it is by studying Pushkin that Markov has developed his mathematical and probabilistic model that is widely used in data analysis and computer science nowadays not only to study language or to develop language models but also to deal with many other general fields such as machine learning tasks and computer vision. Mathematical linguistics afterward participated with a significant growth in the field of language study by providing many models to work on the language (Chomsky, Harris Montague, etc.). Very sophisticated and deep mathematical and computer science theory have been partly the results of interaction between formal linguistics and mathematics (automata theory, lambda calculus, etc.)

[4] More commonly referred to as "statistique textuelle" in the francophone literature

2.2.2. Computational Text Analysis

Textometry have gained a considerable development especially from the late 1950s thanks to computers that allowed computerizing the statistical analysis and performing data management and analysis work including textual data at a non-negligible scale comparing to what could be handled by human agents at that time. If we focus on the development of textometry in the francophone community, we notice very interesting events. At that time, The Besançon Centre of French Vocabulary Study[5] had begun transforming the works of the famous 17th century French writer Corneille into a computer-supported format. After that, the project of the treasury of the French Language[6] started in Strasbourg. These new developments encouraged Charles Müller to start working on the analysis of the textual data of Corneille's works. Lexical-based textual analysis was born (Müller 1967). Few years later, another paradigm which breaks away from the assumption of the linguistic norm universality and uniqueness of the standard, and led by linguistic statisticians such as Jean-Paul Benzécri and then André Salem, emerged by emphasizing the linguistic concern for the systematic analysis of texts, and by adopting multivariate and multidimensional analysis to achieve this quest. (see (Beaudouin 2000) for more details about the nexus between the lexical-based and the linguistic-based textual analysis).

In its early years, textometry was limited to the analysis of stylistic aspects of literary texts, but then it was generalized to texts related to other subjects such as political texts in the 80s, socioeconomically related texts in 90s, and the web in the 2000s. Textometry has been interacting with many other fields throughout its development such as natural language processing, statistical computing, and artificial intelligence. In some sense it has evolved to what is known as computational text analysis.

Computational text analysis is the research field interested in the discovery and measurement of prevalent attitudes, concepts or events in textual data (O'Connor 2014). It kept interest in studying problems related to style and literary subjects throughout its development as well. Ramsay (2007) emphasizes this point by saying that: "*computational text analysis has been used to study problems related to style and authorship for nearly sixty years. As the field has matured, it has incorporated elements of some of the most advanced forms of technical endeavor, including natural language processing, statistical computing, corpus linguistics, and artificial intel-ligence. It is easily the most quantitative approach to the study of literature, the oldest form of digital literary study, and, in the opinion of many, the most scientific form of literary investigation*".

2.2.3. Introduction to Computational Stylistics

Computational text analysis, which is concerned and interested in studying subjects related to style, is commonly known nowadays as computational stylistics[7]. Computational stylistics is a subdomain of computational linguistics located at the intersection between several research areas such as natural language processing, literary analysis, stylistics and data mining. The goal of computational stylistics is to extract style patterns characterizing a particular type of texts using computational and automatic methods. In other words, it aims to investigate texts from the

[5] "Le Centre d'Étude du Vocabulaire Français de Besançon"

[6] "Trésor de la Langue Française"

[7] Computational stylistics still can be sometimes referred to (simply) as 'Stylometry' in the literature as well

standpoint of individual style (style related specifically to a certain author) or functional style (style related to more complex concerns or subjects such as genres or registers) in order to find patterns in language that are not or very hardly demonstrable without computational methods and linked to the processes of writing style in its wider sense (Craig 2004).

Concerning the individual style analysis, when investigating the writing style of a particular author, the task will be to automatically explore linguistic forms of his style, which includes not only distinguishing features but also deliberate overuses of certain structures compared to a certain linguistic norm[8] (Mahlberg 2013). However, as in the general context, the notion of style in the context of computational stylistics appears to be wide enough, and is manifested on several linguistic levels: lexicon, syntax, semantics and pragmatics. Each level has its own markers of style and its own linguistic units that characterize it and are subsequently interesting to be investigated. Computational stylistics is part of the much wider field of digital humanities that covers a range of methods and approaches intersecting both computing and disciplines of the humanities (Siemens and Schreibman 2013).

As it addresses questions of style, computational stylistics shares many commonalities with one of its very successful application, that is computational authorship attribution (by opposition to traditional authorship attribution) in which one assigns a text of unknown authorship to one of some candidate authors based on the stylistic information extracted from documents written by them. However, rather than concentrating on those subconscious traits that may constitute an author's fingerprint, computational stylistics seeks to study those features of an author's style that are not only distinctive but also intentionally used by the author. Computational stylistics is linked to many other related tasks such as stylistic-based text generation (Hovy 1990), and automatic readability and complexity assessment (Pitler & Nenkova 2008).

Recently, many other varieties of applications for computational stylistics have emerged especially with the democratization of internet and the availability of huge amount of machine readable text collections which made the problem of managing these large text collections increasingly both crucial and important, especially from a forensic standpoint. Computational stylistics has been shown to be valuable for the development of information management and retrieval systems to handle such problems. For instance, it is very useful for filtering web-document based on their appropriateness, or for detecting abusive or threatening messages. Various general applications may vary from organizing and retrieving documents based on their writing style or quality to performing some stylistic text classification (Kessler et al. 1997).

However, the fields of research in which the notion of style find its strong arguments are the computer-assisted literary analysis and computer-based literary criticism. As mentioned before, computational stylistics techniques have been used for nearly sixty years to study literary analysis questions relating style (see (Siemens & Schreibman 2013) for a more detailed discussion and overview of this point). As we will explain it later in this chapter, first works focused more on lexical traits such as word counts. Later on, more complex stylistic traits have been taken into account.

[8] Not necessarily considered to be a universal linguistics norm as done in the early stages of lexical-based textual analysis

2.2.4. Relationship between Computational Stylistics and Literary Analysis

When talking about computational stylistics and the notion of style in literary texts, a very important and interesting question comes into mind. It concerns the relationship between these two subjects. Is the computational stylistics meant to replace the traditional literary stylistics? If yes, how does it do that? If no, what can computational stylistics give to the traditional literary stylistics and interpretation, and what is the nature of the interaction between them?

Well, as we can conclude from the reading and from the work done so far in the field of computational stylistics, we can claim that the answer is no. Nevertheless, there exist many strong contributions that computational stylistics makes to traditional literary stylistics and interpretation. In fact computational stylistics can be seen as an assistant generating improvement for the literary analysis and interpretation process (see Figure 2). Actually, the methodology commonly used in culture sciences in general and in literary studies in particular, are based on research activities that are in most cases more dominated by intuition developed through reading literary texts (Ganascia 2015) and hand-crafted features and manipulations.

As pointed out by Mahlberg (2013) in her book dedicated for the computer-assisted stylistic study of Dickens's fiction corpus, among the obvious contributions that such methodology can bring to literary stylistics is its potential to add systematicity and objectivity to the process of an analysis by providing quantitative data in a systematic and objective way for a given phenomenon under investigation. The author gives the simple yet illustrative example of a concordance analysis tool that can, for instance, help tracing linguistic features exhaustively throughout a whole text.

Another aspect that can be considered as a valuable contribution of computational stylistics to the literary analysis studies is the algorithmic aspect. In fact, from the computational point of view, computational stylistic methods are framed as algorithms that are able to extract, count and rank linguistic features in a given text based on measures of *interestingness*.

A debate is currently on-going on whether computational stylistic methods should be a way to make literary criticism more scientific. The influential book by Ramsay (2011), *Reading Machines, Toward an Algorithmic Criticism*, in which he discusses the viability of computational and data-driven techniques for literary criticism, tells us that it may.

The present stage of computational stylistics may not yet allow computational tools to be applied to discover completely new facts about literary style. At the present moment, such tools can at most constitute an aid for literary critics to automatically substantiate known facts. More specifically it is in fact the investigation of already known facts that can help computer scientists and computational linguistics specialists to fine-tune their methods. Nevertheless, specialists can find confirmation of known facts and thus substantiate their claims with more data.

Finally, these methods can be applied to that part of literature that Franco Moretti (2005) calls the archive (as opposed to the canon), notably to works whose lower literary prestige and high number make computational methods more attractive. This goes precisely in line with extending the coverage scope of stylistics analysis not only to include more texts considered to be less interesting, but also to generalize the analysis process to more linguistic features somehow as an extensive exploratory search way. And this is what Craig (2004) describe as the alternative approach: "*This follows the lines of a more traditional, hypothesis-driven design. The alternative approach is through exploratory data analysis, in which the researcher changes all possible*

parameters in the search for a revealing finding. Performed with due cautions, this may lead to discoveries that might be obscured by the starting conditions of a more fixed study. As the cost in time of collecting data and manipulating it and presenting it visually has come down, the attractiveness of exploratory analysis has increased".

While developing those arguments for the cause of computational stylistics, it is actually necessary to emphasize the fact that computational stylistics does not deny in any case neither the need for the close engagement with the literary text in question (close reading as opposed to distant reading), nor the need for the valuable knowledge and interpretation of the literary researchers and specialists.

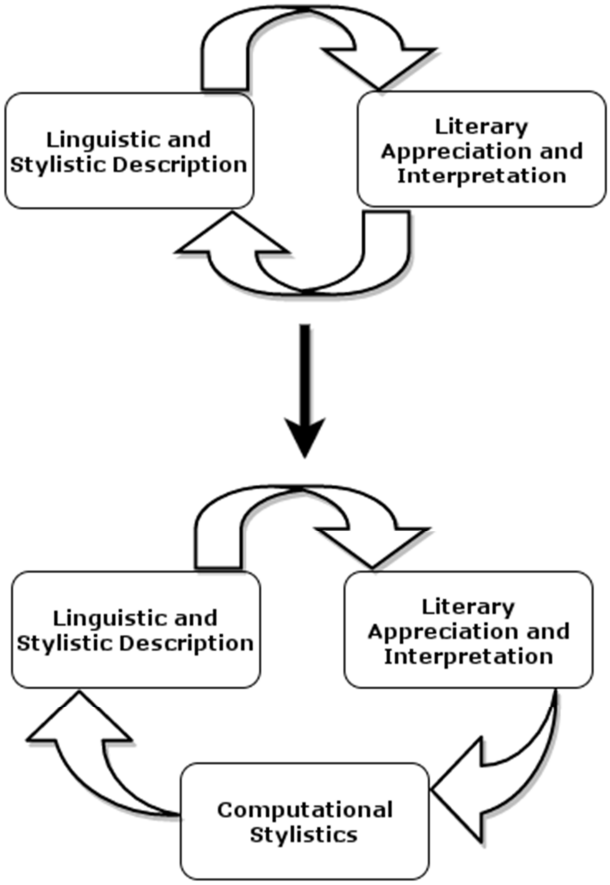

Figure 2. Relationship between computational stylistics and literary analysis and interpretation

2.2.5. Challenges Facing Computational Stylistics

Despite the fact that computational stylistics techniques have been used for nearly sixty years to study question related to literary analysis, and hundreds of works have been published dealing with the notion of style, this domain still struggle with many theoretical and practical issues.

These issues constitute important challenges that computational stylistics (and traditional stylistics as well in some cases) has to deal with in order to strengthen its position as a widespread recognized study field. The issues that computational stylistics is facing can be categorized in three different types: theoretical, practical and structural (community-related) issues.

In the theoretical part, the main issue is related to the notion of style itself. The question to ask here is whether style can be captured. As we have defined it in the beginning of this chapter, many stylisticians think and claim that someone writing's style can be captured either by doing some manual or automatic analysis and reading of his/her written texts. This is actually the very basis of all the fields related to stylistics. However, some other researchers are reserved regarding to this idea. For instance Bordas (2008) claims that style is simply as abstract as an idea: "*Le style est une idée – ce qui ne veut en rien dire qu'il n'existe pas ; il n'y a pas de style, le style, un style, il n'y a que des idées de style, des idées qui sont souvent des imaginations, des projections, voire des fantasmes*" [9].

The main illustration to what have been said regarding this issue is the theory presented by Stanley Fish. Fish claims that the meaning of the stylistics features of a text are only created as it is read and this meaning does not resides within the text. When talking about this issue, Craig (2004) reports that Fish "*argues that the formal features described by stylisticians are meaningless except in relation to the reader's perception of them within a reading situation. When abstracted from this setting they refer to nothing but themselves and so any further analysis of patterns within their use, or comparison with the use of others or in other texts, or relating of them to meaning, is entirely pointless*".

In the practical part, the main issue is related to the quantitative and computational methods used to both represent and analyze texts. The main question related to this part is whether most important dimensions of individuality and stylistic traits can be captured in a set of quantitative aspects and frequencies.

Many linguistics and literary researchers have been very enthusiastic about the application of automatic and computational methods to the humanities. Some of them such as Johnstone (1996) argues for the value of the quantitative over the qualitative, the methods of tendencies and interpretation over those of rules and instances for both small and large amounts of data.

However, many other researchers are not that much sympathetic to computational and quantitative works. In that matter, Fish (1979) pointed out, in the early decades of computational and quantitative linguistics and stylistics, that the leap from frequencies to meanings must always be a risky one. The interpreter, who attempts to speculate about the world-view or psychology of a writer based on quantitative findings, presents an easy target for dismissive critique as he said.

It is fair, however, to admit that reducing any text, or any collection of texts to an abstract quantitative and numerical form cannot preserve the totality of its meaning and individual traits that makes it unique and different from others. Thus, the question of the quantitative representation and the frequency interpretation must be taken carefully.

Another challenge that computational stylistics is facing is on the structural side is the relationship between its computational part on the one hand and its stylistics and literary part on the other hand. Actually this not something that characterizes only computational stylistics but it is a common property shared among about all the interdisciplinary study fields. The question here is

[9] "The style is an idea - that is not to say it does not exist; there is no style, the style, a style, there are only ideas of style, ideas that are often imaginations, projections, or even fantasies" [translation provided by the thesis' author]

how to manage the relationship and the overlapping between different disciplines intervening in one single interdisciplinary study.

Rudman (1997) stated this issue in his critic paper when talking about the computational authorship studies (this can apply for computational stylistics in general as well): "*Non-traditional authorship attribution studies bring a unique problem to interdisciplinary studies: who is the authority? who is the experimental spokesman? the group leader? Is it the linguist? the statistician? the computer scientist? the rhetorician? Is it the expert in the field of the questioned work: literature? classics? law? philosophy? religion? economics? What journal or journals do we turn to for an imprimatur or even a nihil obstat. A quick scan of my working bibliography shows that non-traditional authorship attribution studies have been published in well over 76 journals representing 11 major fields – not to mention the 50 or so books, 11 dissertations, and numerous conference proceedings*".

Another challenge that relates also to the relationship between the computational part and the stylistics and literary part is the impact that one can or should have on the other. Hoover (2008) was not that optimistic about that issue as he said that: "*quantitative analysis has not had much impact on traditional literary studies. Its practitioners bear some of the responsibility for this lack of impact because all too often quantitative studies fail to address problems of real literary significance, ignore the subject-specific background, or concentrate too heavily on technology or software*".

We think that this issue should be collaboratively addressed by trying to bridge the gap between the computer scientists on the one side, and linguists, stylisticians and literary researchers on the other side. More concretely, this may be achieved by enrolling in collaborative work and joining efforts on common projects that allow everyone to benefit from the knowledge and the working methods of the others.

2.3. Shared Ground and Related Fields

Computational stylistics is an interdisciplinary domain witch shares a common ground with many other disciplines besides literary stylistics and criticism. While it is not possible to cite or define every one of them in this document, it is nonetheless important to explain at least the three disciplines that we consider to be strongly both relevant and related to our study. As illustrated in Figure 3, these three disciplines are corpus linguistics, corpus stylistics and computational linguistics in its broadest sense in which we arguably include Natural Language Processing (NLP).

Knowing that these research domains extensively borrow or share concepts, methods and techniques from one another, it is important to mention that the definitions and explanations given bellow for each discipline are very simplified. They are not meant to be exhaustive in any cases. However, they can serve to highlight existing differences between them.

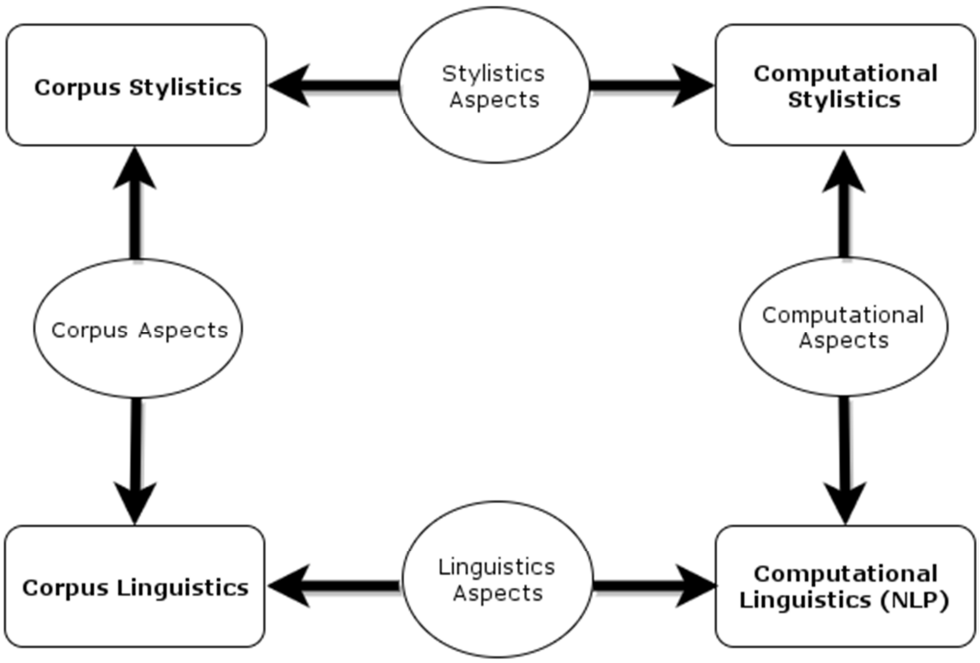

Figure 3. Shared ground and common aspects among computational stylistics and other fields

2.3.1. Corpus Linguistics and Corpus Stylistics

The rapid adoption of computers in the last decades and the on-going expansion of the Word Wide Web in recent years have resulted in the creation of large volumes of text in electronic form including large corpus of literary texts. Consequently, many annotation tools and standards[10] have been developed in order to make modification on these computer-readable texts and to annotate them with additional information and enrichments.

These new resources have contributed to a huge progress in field of the language study and have changed research practices. The field related to the study of style has not been an exception. The availability of electronic and annotated corpus has renewed the research methods used to deal with the notion of style in language study. These resources have considerably helped researchers to stretch the scope of their analysis and to take a new and broader perspective to study the style. New research fields, such as corpus linguistics and corpus stylistics[11], have emerged or distinguished themselves based on these novel developments made both on working resources and working practices.

[10] TEI (Text Encoding Initiative) for instance is developed by a consortium that has as objective the establishment of "standard norm" for digital texts reproduction and encoding. For details about this standard, see http://www.tei-c.org/index.xml

[11] Although not discussed in this report, computational philology could be seen as one of those fields as well. See (Weisser 2006) for more information about it

Talking about the relationship between the corpus and the style, one can notice how close these two concepts have been considered Magri-Mourgues (2006), in her paper entitled *Corpus et Stylistique,* emphasizes on this point by stating that: "*L'interdépendance entre style et corpus est telle que l'on ne peut se définir sans l'autre, ou plutôt que la définition de l'un entraine corollairement une evolution de l'autre. La variation du corpus d'étude induit des pratiques stylistiques différentes. Le corpus est un objet empirique et structuré selon les enjeux et les objectifs de la recherche. Il est par conséquent toujours contingent, déterminé par l'application que l'on veut en faire*"[12].

In fact, the notion of corpus have been deeply discussed in the literature taking it from its very simplistic and shallow definition of being just a collection of language texts to a more epistemologically profound definition. For instance, Sinclair (1996), when defining the corpus, focuses more on the representatively constraint that should characterize any collection of "pieces of language" in order to be considered as a corpus: "*A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as sample of the language. Note that the non-committal word 'pieces' is used above, and not 'texts'. This is because of question of sampling techniques used. If samples are to be all the same size, then they cannot be texts. Most of them will be fragments of texts, arbitrarily detached from their contents*".

In another point of view, Rastier (2011) criticises the representatively constraint set by linguists such as Sinclair and gives a more goal-directed definition for corpus including both the annotation and the homogeneity aspects: "*Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés: (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications"* [13].

Although the terms "corpus linguistics" and "corpus stylistics" have been used only relatively recently, like computational stylistics, these fields of research and their applications on the literary texts have a long tradition as well. Corpus linguistics is concerned with the study of language on the basis of collections of computer-readable texts and it has been practiced as soon as these texts were available. If we focus particularly on corpus stylistics, we can notice the clear overlapping that it has with computational stylistics both in terms of research objectives and means used to achieve them. But, we can also notice at least one clearly simple yet very important divergence point that makes all the difference between those two disciplines. This difference reside in the focus on which the attention is made when researchers in both sides practice the two disciplines.

Actually, Mahlberg (2013) defines corpus stylistics as the field "*concerned with the application of corpus methods to the analysis of literary texts by relating linguistic description with literary appreciation. [...]Thus, it is an area that combines (at least) two disciplines—taking account of methods and theories from literary stylistics and corpus linguistics alike. This mutual relationship is reflected by the attention that corpus stylistics has starred to receive from both corpus linguists and stylisticians*".

[12] "The interdependence between style and corpus is such that one cannot be defined without the other, or rather the definition of one causes corollary an evolution in the other. The variation in the corpus under study induces different stylistic practices. The corpus is an empirical object and structured according to the issues and objectives of the research. It is therefore always contingent, determined by the application one want to do with" [translation provided by the thesis' author]

[13] "A corpus is a structured grouping of full texts, documented, eventually enriched by annotations and gathered: (i) in a theoretical reflexive way by taking into account discourses and genres (ii) in a practical way for the purpose of a range of application" [translation provided by the thesis' author]

Indeed, corpus stylistics explores the relationship between linguistic units and the contributions they may make to the effects that texts have on readers, its power are most obvious when explicit links between linguistics units and literary analysis are sought (Mahlberg 2013). While on the other side, computational stylistics explores and focuses on the computational and automatic methods used to relevantly extract both those linguistics units and the existing links between them.

From the point of view explained above, and considering that computational stylistics cannot function without a literary interpretation and analysis backing it up, we can strongly stress the fact that corpus stylistics is not only an overlapping field to computation stylistics but a complementary one as well.

2.3.2. Computational Linguistics

In the last decade, there has been a huge progress in the field of artificial intelligence and machine learning, computational linguistics and natural language processing in particular have beneficiated from this progress. In fact, thanks mainly to statistical machine learning, language processing tools have attained levels where fairly accurate linguistic analyses of lexical, morphological, and syntactic properties of texts have become feasible. More concretely, natural language processing is the task of analyzing and generating, by computers, natural languages that humans speak, read and write (Bhattacharyya 2012). Ambiguity which is a pervasive phenomenon in natural language is still a major problem confronting natural language processing (Jurafsky & James 2009). One of the goals of natural language analysis is to produce knowledge, extract and analyze some linguistic units from the analyzed text. To do so, this processing makes use of foundational tasks tackling the different level of language complexity such as morphology analysis, lexical and syntactic analysis, or pragmatics and discourse processing for instance.

Table 1 below summarizes the different linguistic levels and their respective linguistics units that can be subject to a natural language processing task.

Natural language processing can be for sure seen as a support for stylistic analysis. Simpson (2004), in his introductory textbook on stylistics, gives an overview of various levels of language and explained how these basic levels of language can be identified in the stylistic analysis of text. In what follows, we give a very brief description of some of these linguistic levels and we highlight some natural language processing tasks associated with each one of them.

Table 1. Linguistic levels of abstraction and their characterizing units

Linguistic level	**Units**
Phonology/Phonetics	Words' sounds
Morphology	Words' forms
Lexicon	Words' storage and associated knowledge
Syntax	Phrases and sentences' structures
Semantics	Meaning of words and sentences
Pragmatics	Discourse units and text connections

2.3.2.1. Morphology

In this level, the interest is based on how words are formed from their roots through processes like inflexion, derivation, back formation. Obvious processing tasks that can be imagined in such level are stemming and lemmatization, which consist respectively and basically in reducing words to their written root form called stem, or to the base form called lemma that one might look up in a dictionary to know more about the word. Table 2 illustrates the results of both stemming and lemmatization process made on the French sentence:

"`Un silence profond régna soudainement dans la maison !`".

Table 2. Results of morpho-syntactic analysis of a French sentence

Sentence	**Un silence profond régna soudainement dans la maison !**								
Tokenization	Un	silence	profond	régna	soudainement	dans	la	maison	!
Stemming	un	silenc	profond	regn	soudain	dans	la	maison	!
Lemmatization	un	silence	profond	régner	soudainement	dans	le	maison	!
POS tagging	DET : ART	NOM	ADJ	VER	ADV	PRP	DET: ART	NOM	SENT

2.3.2.2. Lexicon

This level deals with vocabulary of language and the knowledge associated with it. The most important task in this level is namely word sense or lexical disambiguation. It refers to the identifcation of the meaning of an ambiguous word depending on the context.

2.3.2.3. Syntax

Syntactic analysis is one of language technology applications that benefited most from the statistical machine learning development. The construction of linguistics resources such as the Syntactic Penn TreeBank (Marcus et al. 1993), combined with machine learning techniques have considerably boosted the language technology domains related to the study of syntax. Part Of Speech tagging (POS tagging) (Cutting et al. 1992) is one of these applications that have reached a fairly accurate performance. POS tagging consists in automatically annotating each syntactic constituent, called token, in a sentence with a syntactic part-of-speech marker (see Table 2 for example of the syntactic analysis produced using such tools). The task of identifying the different tokens constituting a sentence is called tokenization. It generally precedes the POS tagging operation. Many POS taggers have been developed and proposed in the literature to achieve such task for several languages, with more or less the same performance. Some of them have adopted the TreeBank syntactic annotation convention. Others have developed their own annotation

system. TreeTagger[14] (Schmid 1994) is known to be one of the most successful and used syntactic POS tagger in the natural language processing community. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. It can be used to analyze more than 22 languages through specific models. In addition to analyzing the syntactic structure of the input text, TreeTagger gives morphological information (Lemma) for each syntactic constituent. Since we have been using this tool in our thesis, we illustrate in Table 3 its syntactic annotation tags and their significations both in English and in French.

Less accurate in terms of performance than POS tagging, syntactic parsing or deep syntactic processing (Traxler 2014) refers to identifying the syntactic hierarchical structure behind a segment of words. Figure 4 illustrates the hierarchical syntactic structure called parsing tree resulting from parsing the sentence taken as example above.

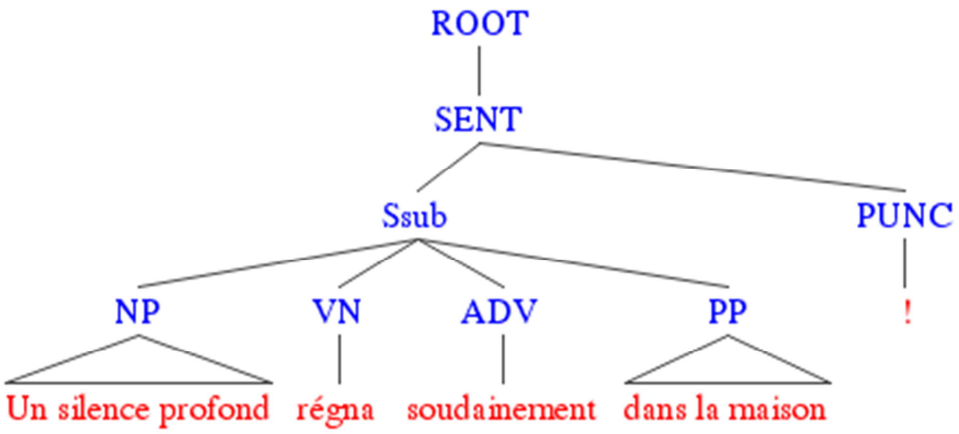

Figure 4. An illustration of a parsing tree, a syntactic hierarchical structure of a textual segment

2.3.2.4. Semantics and Pragmatics

The logical step after analysing the word forms and their structures is to analyse their semantic function in their containing sentence. This is actually a crucial task for someone who wants to access the full meaning of the propositional content of some textual segment. In order to face semantic ambiguity that can manifest itself in the text represented in its original form, many unambiguous formalism and representation have been both presented and used to formalize semantics such as semantic net or conceptual structure for instance (Jackendoff & Jackendoff 1992). The predominant natural language task of such linguistic level is semantic role labelling (Gildea & Jurafsky 2002). Basically, this task consists in automatically identifying events and their participants in some propositional content. This means that the algorithm should be able to answer questions like: "Who did what to whom, where and when?".

[14] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

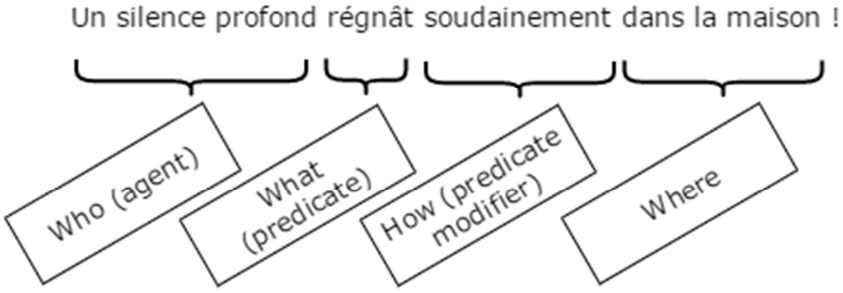

Figure 5. An abstract illustration of semantic role labeling

In the most abstract level, the problem of studying pragmatics involves processing very complex cognitive and linguistics features such as intention and sentiment. It is actually considered to be one of the hardest research domains in computational linguistics. Very little progress has been made so far in automatic processing of pragmatics. It involves processing logical connections among texts as well as connections among text units (clauses, sentences, etc.). One representative task of pragmatic level is shallow discourse parsing which consists in identifying discourse relations between two adjacent or non-adjacent discourse units (Xue et al. 2015).

Table 3. Syntactic annotation tag set and its signification used by TreeTagger

Tag	**Signification**	**Signification (French)**
ABR	Abbreviation	Abréviation (ex. : Dr., M.)
ADJ	Adjective	Adjectif
ADV	Adverb	Adverbe
DET:ART	Article	Déterminant article défini et indéfini (ex. : le, l', un, des)
DET:POS	Possessive pronoun	Déterminant possessif (ex. : ma, ta, son)
INT	Interjection	Interjection (ex. : hé, ah, Hélas)
KON	Conjunction	Conjonction (ex. : que, et, car)
NAM	Proper name	Substantif – noms propres
NOM	Noun	Substantif – noms communs
NUM	Numeral	Nombre – adjectif numéral (ex. : deux, 2016, VI)
PRO	Pronoun	Pronom
PRO:DEM	Demonstrative pronoun	Pronom démonstratif (ex. : celui, ceux)
PRO:IND	Indefinite pronoun	Pronom indéfini (ex. : quelqu'un, tout)
PRO:PER	Personal pronoun	Pronom personnel (ex. : tu, nous, j')
PRO:POS	Possessive pronoun	Pronom possessif (ex. : mien, tien, nôtre)
PRO:REL	Relative pronoun	Pronom relatif (ex. : qui, auquel, dont, où)
PRP	Preposition	Préposition
PRP:det	Preposition plus article	Déterminant contracté (au, du, aux, des)
PUN	Punctuation	Marque de ponctuation
PUN:cit	Punctuation citation	Marque de ponctuation de citation
SENT	Sentence tag	Marque de fin de phrase ("., ?, !")
SYM	Symbol	Symbole
VER:cond	Verb conditional	Verbe au conditionnel
VER:futu	Verb futur	Verbe au futur de l'indicatif
VER:impe	Verb imperative	Verbe à l'impératif de l'indicatif
VER:impf	Verb imperfect	Verbe à l'imparfait de l'indicatif
VER:infi	Verb infinitive	Verbe à l'infinitif
VER:pper	Verb past participle	Participe passé
VER:ppre	Verb present participle	Participe présent
VER:pres	Verb present	Verbe au présent de l'indicatif
VER:simp	Verb simple past	Verbe au passé simple
VER:subi	Verb subjunctive imperfect	Verbe au subjonctif imparfait
VER:subp	Verb subjunctive present	Verbe au subjonctif présent

2.4. Different Approaches to Computational Stylistics

In the last decades, hundreds of works have been done on the field of computational (and quantitative and empirical to be more inclusive) stylistics on different languages and based on both technical and literary different standpoints. Thus, it is extremely difficult to identify some driving lines explaining what differs and what is common among them. However, it is in fact possible to draw some high level categorizations.

2.4.1. Classification Approaches vs. Hermeneutic Approaches

From an abstractive point of view, two different types of approaches have emerged in the field of computational stylistics:

- *Classification approach*, that can be simplified as such: an a priori classification is found in literature (such as Shakespeare's comedies vs tragedies for instance); some relevant linguistic features are identified and counted (such as function words) and finally classification or clustering techniques are used to see whether the a priori distinction holds or not (Craig 2004).
- *Hermeneutic approach*, in which texts are analyzed in order to automatically extract significant features that may later be used by domain experts to produce a better informed and data driven critical analysis of texts (Ramsay 2011). We arguably qualify this second approach as hermeneutic in order to be more inclusive. This includes all the computational stylistic work based on some stylistic interpretation theory without consideing the a priori classification as a mean of analysis. Technically speaking, this approach could be qualified as inductive as well, in the sense that the stylistic analysis could be possibly concerned with the generation of new theories from the analyzed data in such approach.

The classification approach is objectively the one chosen to deal with the stylistics analysis related to authorship studies. More recently, problems deriving from the authorship studies such as authorship attribution and author profiling have gained greater importance due to new applications in forensic analysis, humanities scholarship and some commercial applications, and also due to the development of statistical and the computational methods for addressing the problem. In this thesis, as we have made contributions to the field of computational authorship attribution (see Appendix B and Appendix C), we decided to discuss computational authorship studies more thoroughly in a separate (next) section.

In general, the main advantage of classification approaches over hermeneutic approaches is that they incorporate within themselves a baseline of evaluation in the sense that at least for an abstract goal-directed assessment, one can say that the most relevant stylistic features are the ones capable of reproducing the classification considered in the very beginning and in which the approach was founded.
In other words, such approaches are evaluating the importance of a given stylistic feature based on its predictive power. Moreover, from a statistical point of view, there exist many frameworks

and measures to formally and accurately handle such evaluation. Hermeneutic approaches do not have such advantage, since they are not explicitly relying on an a priori classification. That makes the evaluation of stylistic features extracted from methods based on this approach much harder.

However, if we look to this issue from a different perspective, one should question the assumption that the relevancy of a stylistic feature (its ability to describe the style of a particular text) is related to its predictive power in holding true the a priori classification (its ability to distinct the text in question from different ones).

Actually, at least for the context of analyzing the style of texts written by different authors, which brings us to the authorship studies, this assumption is not always true. This issue have been partly addressed in the critic papers of Craig (1999).

It turns out that the stylistic features that are best capable of identifying the authorship of a given text (stylistic features with high predictive or distinctiveness power) are the ones that act at the abstract level of language, which make them by the way not fully perceptible or consciously controlled. What such features have as property as well, is that they are difficult to be linguistically or literarily interpreted and thus they are not that much interesting from a stylistic point of view. Subsequently, the produced stylistic characterizations that are based on those features do not reflect the important and the relevant stylistic choices taken by authors in question.

Given the facts explained above, we claim that hermeneutic approaches are best suited for dealing with the notion of style and for extracting stylistic features capable of describing the stylistic aspects of a particular type of text without losing the very important point of interpretability, especially if we are looking for studying the style of a particular author in a relatively high level.

2.4.2. Corps-Driven vs. Corpus-Based Methodologies

Hermeneutic approach in its turn can include many methodologies that can be followed to accomplish a stylistic research activity. We can arguably[15] distinguish two main methodologies in that matter: the *corpus-based* methodology and the *corpus-driven* methodology.

A corpus-based methodology considers, as a strong assumption, the existence of some linguistic (stylistic) theories and uses corpora analysis as a mean to test and validate them. In contrast, a corpus-driven methodology derives linguistic (stylistic) interpretations and models on the basis of patterns that are discovered from the analyzed corpus. In other word, it tries to extract and discover interesting patterns that will be subject to a literary interpretation and appreciation in order to enhance the knowledge known about the analyzed corpus. Therefore, the focus in the corpus-based line is made on the stylistic model, while in the corpus-driven line the focus is made on the data (the analyzed literary texts). This fundamental difference will generate two different research processes that can be modeled as ordered methodological steps as follow:

Corpus-based methodology:

1. Assume the existence of a certain stylistic theory

[15] From another perspective, one can suggest that this constitutes a complete different way of categorization in its own, and does not/should not necessarily need to be considered as a subcategorization of hermeneutic approach

2. Design and build an appropriate corpus
3. Quantitatively and qualitatively analyze the corpus
4. Analyze the application of the stylistic theory and eventually confirm it

Corps-driven methodology:

1. Design and build an appropriate corpus
2. Annotate and analyze the corpus both qualitatively and quantitatively
3. Identify the relevant stylistic traits resulting from the analysis
4. Interpret the results and induce some stylistic knowledge

A more thoroughly discussion about the distinction between corpus-based and corpus-driven can be found in (Tognini-Bonelli 2001). Moreover, if we chose to use more technical words to describe the nexus between corpus-based and the corpus-driven methodologies, we can say that the former derives from the deductive paradigm in the sense that it produces some conclusion based on the validation of an a priori theory, while the latter derives from the inductive paradigm, in the sense that it tries to explore propositions from specific observations. Magri-Mourgues (2006) highlights this idea when talking about how corpus could be taken as a study object for a stylistic research: "*Le corpus est un objet empirique et structuré selon les enjeux et les objectifs de la recherche. Il est par conséquent toujours contingent, déterminé par l'application que l'on veut en faire. La constitution du corpus d'étude en stylistique est confrontée à la même problématique que n'importe quel autre corpus d'étude; le chercheur oscille entre deux tendances complémen-taires: une démarche déductive lorsqu'une thèse préalable préside à l'établissement de ce corpus et une démarche inductive quand c'est l'observation de spécificités langagières qui sous-tend l'élaboration d'une théorie*"[16].

Of course, based on these two methodologies, some hybrid methodology can be imagined. In fact Rayson (2008) presented a hybrid one biased more toward the corpus-driven point of view. He describes it as *data-driven* combining elements from both corpus-based and corpus-driven methodologies. It goes as follow:

1. Build: corpus design and compilation
2. Annotate: manual or automatic analysis of the corpus
3. Retrieve: quantitative and qualitative analyses of the corpus
4. Question: devise a research question or model (iteration back to Step 3)
5. Interpret: manual interpretation of the results or confirmation of the accuracy of the model

Finally, it is worth pointing out that the corpus-driven (and data-driven) methodology, even if it gives priority to the data, does not deny the need for some stylistic theoretical assumption in the

[16] "The corpus is an empirical object and structured according to the issues and objectives of the research. It is therefore always contingent, determined by the application one want to do with. The constitution of the study corpus in stylistics is facing the same problematic as any other study corpus; the researcher oscillates between two complementary trends: a deductive approach where a prior thesis governs the compilation of this corpus, and an inductive approach where it is the observation of linguistic specificities that underlies the development of a theory" [translation provided by the thesis' author]

research process, especially when it comes to interpreting from a literary point of view the produced results and appreciating their quality. Mahlberg (2005) explained this fact by saying that: "*Theoretical assumptions cannot be avoided and working in a corpus-driven way does not mean we pretend to do without theory. What a corpus-driven approach aims for, however, is to keep the assumptions minimal*".

2.5. Review of the Authorship Attribution Problem

The computational studies of style have converged into two main overlapping branches of research: computational stylistics on the one hand, and computational authorship attribution on the other hand. The similarity of the computational used methods notwithstanding, the purpose of computational stylistics is different from that of authorship attribution. Indeed attribution methods aim to identify unconscious traits in the work of a given author, which tell him away, and for this reason, are normally defined as fingerprints. Basic features (such as word or sentence length), together with function words distribution, have proved to be very efficient fingerprints. It is imaginable that such traits persist in a single author somewhat independently of the kind of text he is writing; even outside its literary production in a strict sense.

On the other hand literary style is something that the author masters in a more conscious way. It is imaginable that different works of the same author may show different stylistic traits, although others may be found in all of his works.

Moreover, authorship attribution can be clearly framed as a classification problem (who is the most likely author of a text given a set of candidates) and indeed it is applied as such not only to literature but also in forensics. On contrast, computational stylistics is an open ended problem that consists in identifying such traits that are most distinctive of a set of texts, with respect to other ones, as Craig (2004) explained: "*Stylistic analysis is open-ended and exploratory. ... Authorship studies aim at yes or no resolutions...Yet stylistic analysis needs finally to pass the same tests of rigor, repeatability, and impartiality asauthorship analysis if it is to offer new knowledge*".

2.5.1. Problem Statement

Authorship attribution is the task of identifying the author of a given document. The authorship attribution problem can typically be formulated as follows: given a set of candidate authors for whom samples of written text are available, the task is to assign a text of unknown authorship to one of these candidate authors (Stamatatos 2009). This problem has been addressed mainly as a problem of multi-class discrimination, or as a text categorization task (Sebastiani 2002). Text categorization is a useful way to organize large document collection. Authorship attribution, as a subtask of text categorization, assumes that the categorization scheme is based on the authorial information extracted from the documents.

Authorship attribution is a relatively old research field. A first scientific approach to the problem was proposed in the late 19th century in the work of Mendenhall in 1887, who studied the

authorship of texts attributed to Bacon, Marlowe and Shakespeare. More recently, the problem of authorship attribution gained greater importance due to new applications in forensic analysis and humanities scholarship (Stamatatos 2009).

Authorship attribution and stylometry have always been closely related research fields. In fact, authorship analysis relies on the notion of style and on the process of drawing conclusions about authorship information of a document, by analyzing and extracting the stylistic characteristics. This assumes that the author of that document has a specific style by which he can completely or partly be distinguished from another author.

Following this idea, current authorship attribution methods have two key steps: (1) an indexing step based on style markers is performed on the text using some natural language processing techniques such as such as tagging, parsing, and morphological analysis; then (2) an identification step is applied using the indexed markers to determine the most likely authorship. An optional features selection step can be employed between these two key steps to determine the most relevant markers. This selection step is done by performing some statistical measures of relevance such as mutual information or Chi-square testing.

The identification step involves using methods that fall mainly into two categories: the first category includes methods that are based on statistical analysis, such as principle component analysis (Burrows 2002) or linear discriminant analysis (Stamatatos et al. 2001); the second category includes machine learning techniques, such as simple Markov chain (Khmelev & Tweedie 2001), Bayesian networks, support vector machines (SVMs) (Koppel & Schler 2004, Diederich et al. 2003) and neural networks (Ramyaa & Rasheed 2004). SVMs, which have been used successfully in text categorization and in other classification tasks, have been shown to be the most effective attribution method (Diederich et al. 2003). This is due to the fact that SVMs are less sensitive to irrelevant features in terms of degradation in accuracy, and permit one to handle high dimensional data instances more efficiently. The typical process of authorship identification is illustrated in Figure 6.

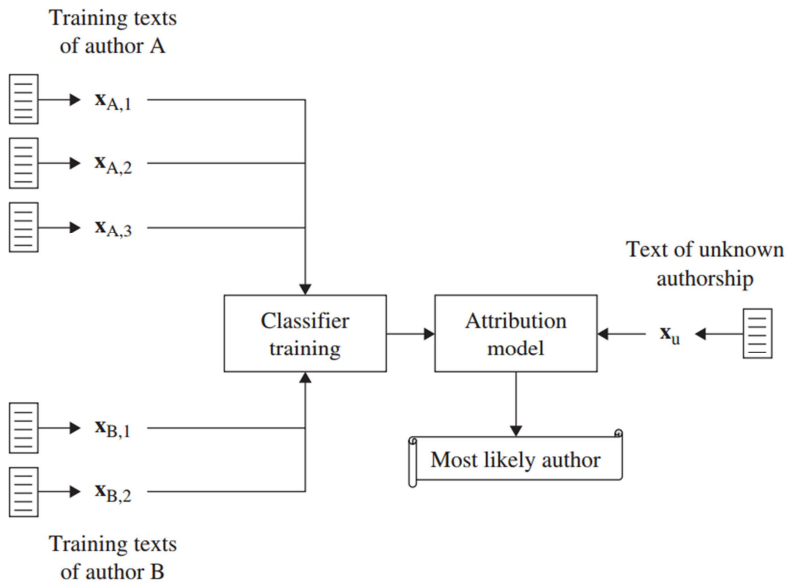

Figure 6. Typical process of authorship attribution (Stamatatos 2009)

2.5.2. Stylistic Features for Authorship Attribution

To achieve high authorship attribution accuracy, one should use features that are most likely to be independent from the topic of the text. Many style markers have been used for this task from early works based on simple features such as sentence length and vocabulary richness (Yule 1944) to more recent and relevant works based on function words (Holmes et al. 2001, Zhao & Zobel 2005), punctuation marks (Baayen et al. 2002), part-of-speech (POS) tags (Kukushkina et al. 2001), parse trees (Gamon 2004) and character-based features (Kešelj et al. 2003). There is an agreement among different researchers that function words are the most reliable indicator of authorship. There are two main reasons for using function words in lieu of other markers. First, because of their high frequency in a written text, function words are very difficult to be consciously controlled, which minimizes the risk of false attribution. The second is that function words, unlike content words, are more independent from the topic or the genre of the text, so one should not expect to find great differences of frequencies across different texts written by the same authors on different topics (Chung & Pennebaker 2007). The POS-based markers are also shown to be very effective because they partly share the advantages of function words (Stamatatos 2009).

Despite the fact that function word-based markers are state-of-the-art, they are basically relying on the *bag of words* assumption, which stipulates that a text is a set of independent tokens. This approach completely ignores the fact that there is a syntactic structure and latent sequential information in the text. DeRoeck et al. (2004) have shown that frequent words, including function words, do not distribute homogeneously over a text. This provides evidence of the fact that the bag of words assumption is invalid. In fact, critiques have been made in the field of authorship attribution claiming that many works are based on invalid assumptions (Rudman 1997) and that researchers are focusing on attribution techniques rather than coming up with new style markers that are more precise and based on less strong assumptions. Table 4 presenting the taxonomy of the authorship analysis gives more information about this field and other related special cases.

Table 4. A taxonomy for authorship analysis (Zheng et al. 2006)

Problems	
Category	**Description**
Authorship Identification	Determines the likelihood of a particular author having produced a piece of writing by examining other writings by that author
Authorship Characterization	Summarizes the characteristics of an author and determines the author profile based on his/her works.
Similarity Detection	Compares multiple pieces of work and determines whether they are produced by a single author without actually identifying the author.
Features	
Category	**Examples**
Lexical	Average word/sentence length / Vocabulary richness
Syntactic	Frequency of function words / Use of punctuation
Structural	Paragraph length/ Indentation Use of a greeting statement /Use of a farewell statement
Content-specific	Frequency of keywords
Techniques	
Category	**Description**
Statistical Analysis	Uses statistical methods for calculating document statistics based on metrics to analyze the characteristics of the author or to examine the similarity between various pieces of work.
Machine Learning	Uses classification methods to predict the author of a piece of work based on a set of metrics.

2.6. Overview of the Analysis of the Syntactic Aspect of Style

Historically, most of the work done in computational stylistics and textual data analysis of literary texts is focused on lexical aspects, especially in the first decades of the discipline. Moreover the few works that deal with the syntactic aspect of style were either rule-based ones, or more focused on syntactic characteristics that can be analyzed without the need for any advanced natural language processing tools such as counting and analyzing function words' use.

Researchers mainly used techniques based on some automatic lexical-based linguistic form's counts in which frequency data are collected from the analyzed texts eventually alongside with other simple stylistic traits such as sentence length or text length. This data is analyzed afterward through statistical assessments for some specific application such as concordances and other word-count applications (Raben 1965, Burrows 1987, Landow 1993). There are many reasons for that. The first obvious reason that one can notice is the low computing power of computers at the time of the emergence of such research domains comparing to what we have now. It is somehow astonishing to notice that the computing power of even the simplest smartphone today would far exceed the computing power of the computer that put the man on the moon in the Apollo mission! Computers in the past neither were democratized as they are nowadays. It waited until the late 80s to see the revolution of the personal computers. It is also worth mentioning that most languages at that time were under-resourced in terms of computer-readable texts.

The second reason that let researchers take a bigger focus on the lexical aspect of the style instead of syntax is the non-availability of fairly accurate syntactic analysis tools such part-of-speech taggers or syntactic parsers. Researchers interested in studying syntax were forced to perform some manual syntactic annotation and analysis which limited considerably the amount of texts that can be put under analysis. With the development of new artificial and machine learning techniques, things have changed in the sense that fairly performing (not perfect though) syntactic tools are being available for many languages, especially for those based on Latin script.

In fact, most of the syntactic tools available today were trained on journalistic texts which make their performance on literary texts, which have their own linguistics particularities comparing to journalist texts, less than the performance reported on some standard testing sets. However, the time gain that can be generated and the advantage of being able to process huge amount of texts by using such tools fairly exceed the inconvenient generated by the error rate that can be expected in an automatic syntactic analysis.

The third reason is more theoretical since it is related to the linguistic property of language as opposed to the two technical reasons presented above. Since syntax is more related to the study of language structure and phrasal hierarchies, it is more abstract. In that sense, syntax differs completely from lexicon and semantics by operating on a more abstract level of linguistics. The meaning of some syntactic structure is very difficult both to define and to imagine. The representation of some lexical items such as "car" or "love" in someone's head is much less abstract the representation that one can have on some syntactic structure for instance. It is thus easier for some reader to have a perception about the words chosen by the author than the order in which they are arranged. It is easier to imagine the meaning of words than their syntactic categories as well. Consequently, it is easier for a stylistic researcher to assess the research work he is carrying on when dealing with lexical aspects.

2.6.1. Approaches to Investigating the Syntactic Style

By focusing on the syntactic point of view of the notion of style, one can notice that the written text in general is a very syntactically regulated phenomenon in the sense that not all the syntactical combinations are allowed to construct a well formed syntactic sequence that can carry a semantic meaning. There are two main factors acting at two different levels that regulate the syntactic order of a text. The first one is the grammar that acts on the phrase level by restricting the syntactic variations via a set of syntactic rules. These syntactic rules forbid certain syntactic sequences that are considered invalid, and allow other valid ones (syntactic sequence that respects these rules). The second element is (roughly speaking) the genre of the text which acts at the sentence level. In fact it is clear that a text written in verses will significantly differ from a text written in prose in terms of the syntactic forms that are incorporated on each one of them. This is due to the linguistic constraints imposed by the rhetoric of the genre. These two elements will introduce a certain statistical order into the syntactic sequences.

From a more practical point of view, one can divide the work done in computational stylistics about the syntactic style into two main lines of research:

- *Paradigmatic line* which is based on the quantification of some simple style descriptors, generally lexemes annotated on the basis of semantic characteristics, words forms, or morpho-syntactic tags and then the generation of some kind of generic properties about the studied text. For example, the work done by Biber (2006) in his analysis of academia's discourse constitute an illustrative prototype of such approach. In his work, Biber focuses on the analysis of the morpho-syntactic tags and highlights the over-employment (overrepresentation) of the first personal pronouns for instance.

The biggest advantage of the paradigmatic approach is its simplicity and the relative intelligibility of the produced results. However, in the minus side, this approach is based on the logic of "texts destruction, analysis then reconstruction". Thus, it neglects the very important latent contextual and sequential information in the text. Such inconvenient makes the results too far apart to be operational.

- *Syntagmatic line*, as opposed to the paradigmatic one, favors the combinatorial analysis methods of the morpho-syntactic units in order to identify the preferred syntagmatic sequences or structures (even if not necessarily continuous) that characterize a particular type of texts.

This approach takes into account the contextual and sequential information which constitute valuable information for the analysis process. However, this additional information comes with costs. In fact, the descriptors resulting from such approach are sometimes relatively difficult to interpret. This is actually the research line that we favor in our thesis. In what follows, we present some linguistic units investigated in such paradigm. In Section 2.7 (Strongly Related Works), we report in more details three works based on this line as well.

2.6.2. Stylistic Features of the Syntagmatic Approach

A variety of linguistics units have been investigated for the analysis of style using syntagmatic approaches. In what follow, we present only the units that we consider to be the important ones. For further details, examples of works dealing with such units and more, please refer to the book *Grammar of Genres and Styles: New approaches* (Charnois et al. 2016).

2.6.2.1. Lexical bundles

These stylistic units are not directly related to syntax. However, they constitute a very good example of units that investigate the order between the words in a text rather than the raw word forms' counts. Known also as multiword lexical chunks, formulaic sequences, lexical phrases or simply n-grams in the literature, they are recurrent groups of words that occur repeatedly together within the same register in a text (Biber et al. 2004). Such units are very useful for studying the style of texts written by non-native speaker of some language.

2.6.2.2. Collocational frameworks

A collocation is basically a sequence of words that occurs more than once in identical form and which is grammatically well-structured (Kjellmer 1987). Collocation frameworks, also known as phrase-frames, are collocation constituted from high-frequency function words as fixed elements incorporating some variable internal lexical gaps. Renouf & Sinclair (2014) have extracted and studied many collocation frameworks in English such as:

- a + ? + of
- be + ? + to
- for + ? + of
- too + ? + to

2.6.2.3. Syntactic patterns

Syntactic patterns[17] (motifs) are recurrent syntactic structures that may possibly combine either, depending on their definition, different levels of linguistics abstraction (word forms, lemmas, POS tags) (Quiniou et al. 2012, Longrée et al. 2008) or just different level of syntactic abstraction (syntactic group, syntactic category, word) (Ganascia 2002), as well as syntactic relationships (Tutin 2009). Therefore, they have both a syntagmatic and a multilevel nature.

They constitute a very promising line of research since they present a fair balance between linguistic and stylistic complexity on the one hand, and the intelligibility of the extracted pattern on the other hand. Renewed interests in such linguistic units have been manifested recently for stylistic applications as well as for other purposes. In our thesis, we rely on such patterns in our quest to describe the stylistic aspect of the text under investigation in its syntactic dimension. In the next section of this chapter, we present in more depth two works dealing with such style descriptors.

[17] Also referred to (more specifically) as "morpho-syntactic patterns", or just as linguistic patterns, depending on their conception. In this dissertation, we rather prefer the term "morpho-syntactic patterns". However, we sometimes use the term "syntactic patterns" or just "patterns" for simplicity.

2.7. Strongly Related Works

Among the several works dealing with computational stylistics in particular and computational linguistics in general (with a special interest in style) that we have studied and read, in this section we report on three strongly related works that we consider to be very important and influential to the work carried out in the thesis. They are interesting both from the theoretical and the practical standpoint. The three selected works are:

- Recurrent Segments and the Statistical Analysis of Text Data (Salem 1986)
- Extraction of Syntactical Patterns from Parsing Trees (Ganascia 2002)
- Discovering Linguistic Patterns using Sequence Mining (Béchet et al. 2012)

The first work is relevant in two main points. The first one is that it presents a practical foundation for the syntagmatic approach that takes into account the sequential information in the text. It breaks away from traditional analysis method based on graphic forms counting that fall under the paradigmatic approach and the bag of words assumption. The second point is that this work incorporates an interestingness measure (even though not explicitly called and presented as so in the paper) to objectively identify the most relevant linguistic patterns (recurrent segments).

As we thoroughly explain in the next chapter, our approach is based on such line of research, that is to say a syntagmatic approach which makes use of some interestingness measures to rank and extract the most interesting stylistic forms.

The second work brings many elements that we have made use of in our work. It tackles the very hard problem of extracting pattern from syntactic trees. This work presents what can be considered as a knowledge discovery process for the extraction of stylistic patterns from syntactic trees. These trees are produced using a deep syntactic analyses applied on classic French texts. Even if the syntactic parsers were not at that time as accurately performing as they are nowadays, this work formalizes the notion of a syntactic pattern that can be extracted from syntactic structures and makes use of data mining as part of its process.

The last work is important as it constitutes a proof of the utility that sequential data mining methods can bring for the linguistics analysis of texts. Based on techniques such as sequential pattern mining, relevant and understandable patterns that characterize specific type of text can be extracted. Indeed, the sequential pattern mining techniques constitute a key element in the knowledge discovery process developed in our thesis. In what follows, we present in more details these three works.

2.7.1. Recurrent Segments and the Statistical Analysis of Text Data

The first work that we start discussing in this section is the one done by André Salem (1986) entitled *Segments répétés et analyse statistique des données textuelle* (Recurrent Segments and the Statistical Analysis of Text Data).
Actually, this is a very interesting work that presented not only a new analysis method and interesting results but also a new approach to textual data analysis. This work is a lexicometric study based on what Salem called a *Recurrent Segments* or more precisely the inventory of

recurrent segments. A segment is basically a series of graphic forms unseparated by strong punctuation that appear more than once in a text corpus.

Salem used a method based on recurrent segments to detect the number of units composed of several elements repeated in the same order in different locations within a corpus. He found out that some of these units reoccur with great frequency.

He proceeds to the analysis of recurrent segments in terms not only of raw frequency but on location as well. Certain segments studied are composed of elements regularly distributed throughout the corpus of the text that must be directly indexed. Furthermore, typologies derived from recurrent segments can be applied directly to the study of the chronologic evolution of a corpus.

Salem starts by motivating his working method, he emphasizes the fact that textual data analysis does not consist only in doing some lexical form counting (lexical inventory) and that it should go further by analyzing the order in which these lexical forms appear, their repetitions and their repartitions on the corpus under analysis, which constitute a valuable sequential and syntagmatic information that should be taken into account in any textual analysis project. He pointed out that this syntagmatic dimension was completely ignored by the lexical-based textual analysis point of view and that it was needed to satisfy the precise demands of discourse analysts.

The corpus chosen by Salem to undergo the study is DUCH96 used as a basis for the work on the discursive configurations of the Jacobin discourse. This corpus consists of 96 issues of the *Pére Duchesne* published during the period between the 13 July 1793 to the end of the publication of the newspaper March 12, 1794. The corpus is divided into 8 more or less equal parts on a monthly basis (from M1 to M8).

The recurrent segments were extracted from this corpus. Figure 7 presents a table containing some of them. As Salem stated, these examples show the benefits that such syntagmatic reading can bring to the study of recurring segments with respect to the simple linear reading based on the lexical inventory. However, he does not miss to note the extremely large number of extracted segments of different lengths, which constitutes a serious obstacle to making the analysis method reliable. Thus he limited his interest to the recurrent segments present at least 5 times.

Freq.	Long.	Segment
41	6	LA GRANDE COLERE DU PERE DUCHESNE
35	3	TONNERRE DE DIEU
29	5	à CHIEN ET à CHAT
26	4	LE COUP DE GRACE
17	5	TOUS LES COUPS DE CHIEN
16	6	LA PLUIE ET LE BEAU TEMPS

Figure 7. Some recurrent segments extracted from the corpus under investigation (Salem 1986)

Furthermore, Salem presents a method to evaluate the interestingness of the extracted recurrent segments in terms of characterizing the analyzed corpus. The method is based on what he called the *recurrent neighborhood* ("Voisinage recurrent"). He specifies that in conjunction with the

statistical methods operating from graphic forms, the analysis of the recurring neighborhoods allows selecting segments whose distribution is of interest statistically without needing to make a prior segmentation into disjoint units.

Figure 8 illustrates the table presenting the recurrent neighborhood of the word form “HOMMES”. In this table, we can notice that unlike the occurrences of the word form “HOMMES” that are rather evenly distribution throughout the corpus; the occurrences of segments containing this word form are not. For instance the segment “HOMMES D’ETAT” are all appearing in the first two parts of the corpus.

	M1	M2	M3	M4	M5	M6	M7	M8	TOT
HOMMES	41	35	27	24	20	16	47	26	236
ETAT	13	4	3	0	1	3	0	1	25
HOMMES D'ÉTAT 236 1275 25	11	3	0	0	0	0	0	0	14
LES HOMMES D'ÉTAT 4747 236 1275 25	4	2	0	0	0	0	0	0	6
DES HOMMES D'ÉTAT 1961 236 1275 25	5	0	0	0	0	0	0	0	5
NOUVEAUX HOMMES D'ETAT 36 236 1275 25	2	1	0	0	0	0	0	0	3

Figure 8. Recurrent neighborhoods of the word form "HOMMES" and their distribution in the corpus (Salem 1986)

Another interesting segment reported in this work based on the same idea as the previous one is “GENERAUX SANS-CULOTTES” (see Figure 9). This example chosen by the author from many others shows the interest there is to directly study the distribution of some units whose existence and importance we would not perhaps suspect by simply studying the isolated forms counts.

The previous example illustrates that the comparative analysis of the distribution of occurrences of the segments with the forms they contain represents a very interesting measure. As pointed out by the author, contrary to what could reveal respectively the relatively balanced distribution of each one of the form “GENERAUX” and “SANS-CULOTTES”, there exists a strong imbalance in the occurrences’ distribution of the “GENERAUX SANS-CULOTTES” segment. Using a concordance, tool one can discover that the “GENERAUX” form, which refers in the first period to noble generals that had to be chased from the army, refers almost exclusively in the second half of the corpus to “GENERAUX SANS-CULOTTES” where it was in this time needed to defend them against all attacks.

The author concludes from this study that it is clear that some units (segments) consisting of two, three or four word forms, such as the one presented above, strongly deserve to be included as such in indexes or concordances.

	M1	M2	M3	M4	M5	M6	M7	M8	TOT
GENERAUX	6	5	4	0	4	11	1	2	33
SANS-CULOTTES	58	57	52	43	45	53	41	52	401
GENERAUX SANS-CULOTTES	0	0	0	0	0	7	1	2	10

Figure 9. Distribution of the two forms "GENERAUX", "SANS-CULOTTES" and the segment "GENERAUX SANS-CULOTTES" in the corpus (Salem 1986)

2.7.2. Extraction of Syntactical Patterns from Parsing Trees

The second interesting related work that we present in this section is the one done by Ganascia (2002), entitled *Extraction of syntactical patterns from parsing trees.*

In his paper, Ganascia presents a method capable of extracting clusters of similar recurrent patterns from any stratified ordered trees. The similarity on which the clustering algorithm is based is the generalized edit distance.

What is more interesting about this work is that the presented method is used for a stylistic text mining purpose. Precisely, the aim was to detect recurrent syntactic patterns in texts drawn from classical literature.

To fully understand the work, it should be more convenient to briefly define the Stratified Ordered Tree (SOT) and how a parsing tree can be seen as one. So basically, a SOT is a labeled tree structure in which two constraints have been applied. First, it should be an ordered tree where left to right order between siblings is significant. More details about this data structure can be found in (Ganascia 2001). The order property is natural in texts which can be explicitly represented as such by noticing that a text basically is a sequence of sentences and each sentence is a sequence of words and punctuation. The second constraint is the existence of a sort function representing the stratification property which once applied to the labeled ordered tree makes sure that the sort of the sons is identical to, or immediately follows that of the father. In the case of parsing tree, the corresponding sort to apply is the following: Text $<$ Sentence $<$ Syntactic group $<$ Category $<$ Word.

Figure 10 summarizes the whole processing chain that transforms a natural language text into a set of frequent patterns.

In the first component of the chain, the system takes as an input a natural language text that is to say a sequence of sentences, then it performs a deep syntactic analysis that associates labels to words/punctuations (noun, verb, etc.) and to groups of words (noun group, verb group, etc.). The performed analysis transforms texts into trees or forests, i.e. into sequences of trees. One limitation of such process is that the result of the analysis has to be structured in a SOT.

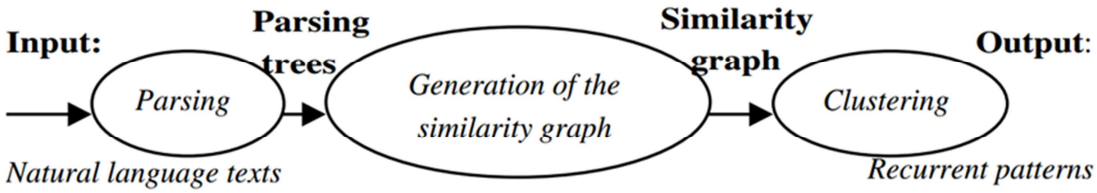

Figure 10. The processing chain (Ganascia 2001)

The second component of the chain builds what Ganascia calls a similarity graph, which is a labeled graph making the distances between patterns explicit when they go above a fixed threshold. As Ganascia points out, the similarity graph, which constitutes the main input of the clustering component and includes all the patterns that generalize sub-trees of the input SOT, is a key-point in the overall method since it generates all general patterns including non-balanced ordered trees. In his paper, Ganascia explains and reports in detail the algorithm used to generate the similarity graph.

Figure 11 illustrates a non-balanced pattern covering the French textual segment "`Elle exécuta ce qu'elle avait projeté :`". Detail about the parsing formalism can found in (Vergne 1999).

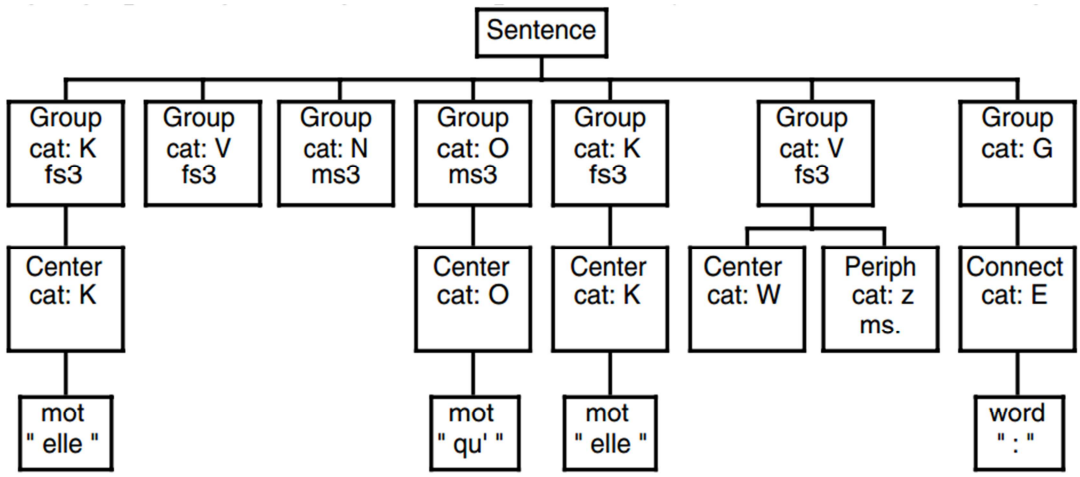

Figure 11. An non-balanced pattern covering the textual segment " Elle exécuta ce qu'elle avait projeté :" (Ganascia 2001)

Finally, to compute the similarity graph, all pairs of patterns T_1 to T_2 have to be produced and then the generalized edit distance is used to compute the value of this edit distance between T_1 and T_2.

Once all similarities between patterns are recorded in the similarity graph, the third and last component of the processing can be applied. It consists of detecting the highly connected sub-graphs of the similarity graph using center-star algorithm. Very briefly, the algorithm just chooses the pattern that maximizes the similarity with other members of the cluster and minimizes the similarity with members of other classes.

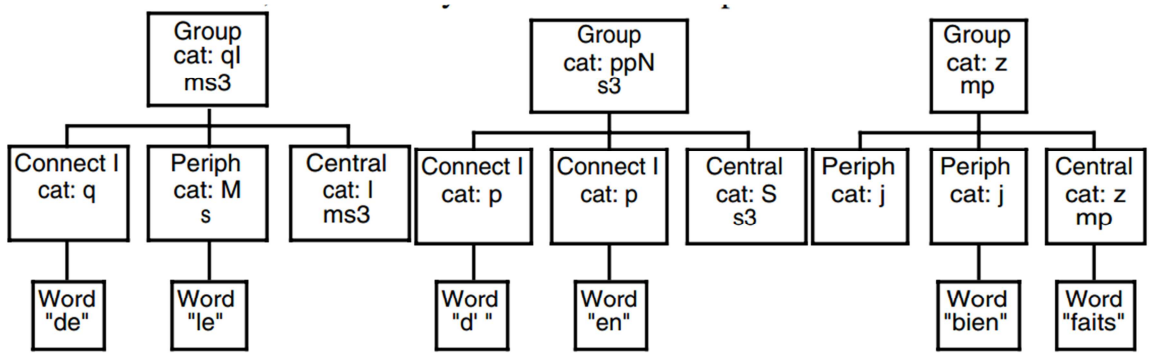

Figure 12. Three patterns present in the Lafayette texts without any occurrences in other texts (Ganascia 2001)

As case study, this processing chain was used to detect recurrent syntactic patterns in classical literature texts written by Madame de Lafayette (two texts, a short story entitled *La comtesse de Tende* and a famous novel, *La princesse de Clèves*). For a comparative purpose, more than 25 short stories from three 19th century authors, Guy de Maupassant, Georges Sand and Marcel Schwob, were used.

In Figure 12, Ganascia reports, among others, three interesting patterns extracted using this method. The patterns seem to be very interesting in describing the syntactic aspect of Madame de Lafayette writings. For instance, the first pattern (on the left) covers among others the following French expression in the studied texts:

- `de le supplier`
- `de l'éviter`
- `de l'aimer`

And others like: "`de la tromper`", "`à le servir`", "`pour l'obliger`".

The second (in the middle) covers among others:

- `d'en avoir`
- `d'en attendre`
- `d'en garantir`
- `sans en avoir`

The third (in the right) covers the following three fragments:

- `admirablement bien faits`

- `parfaitement bien faits`
- `très bien fait`

2.7.3. Discovering Linguistic Patterns using Sequence Mining

In the work entitled *Discovering Linguistic Patterns using Sequence Mining*, Béchet et al. (2012) presents a syntagmatic processing chain from which the approach considered in our thesis was greatly inspired.

The main contribution in our eyes in this paper is the fact that in this syntagmatic-based study, the authors have shown the interest of using sequential data mining methods for the linguistic analysis of large texts. They have shown that relevant patterns that characterize specific type of texts can be extracted using sequential data mining techniques such as sequential pattern mining. They have considered the text as a set of sentences and each sentence as a sequence of ordered syntactic (POS-tag) or lexical (lemma) items. Each item in the sequence corresponds to one token in the sentence respecting the order. Using this configuration as input for the sequential pattern mining algorithm, they point out that their method is better than machine learning methods such as Hidden Markov Models or Conditional Random Fields, in the sense that it produces outputs that are more understandable by humans.

The authors emphasize the point that since their approach is based on sequence mining techniques, it is independent from the language under investigation and its linguistic properties. They state that their work can be adapted to other information extraction applications such as studying relationships between named entities.

They apply their approach to learning linguistic patterns for discovering phrases denoting judgment or sentiment in French texts, and more generally qualification called appositive qualifying phrases. Basically, to do so, they proceeded on a two-step basis.

The first step consists in an extraction task in which patterns expressing information according to different levels of generality are extracted (patterns that combine different levels of abstraction, e.g., words, lemma, part of speech tags). Secondly, because of the high number of extracted sequential patterns, a validation (pattern selection) task is needed. To address this issue, the authors propose a selection tool allowing a user to easily navigate within the pattern space and selectively validate some sequential patterns as interesting ones. The navigation and validation tool (illustrated in Figure 13 which is extracted from the authors' paper) is based on the property of partial order between patterns[18].

To experiment their approach, the authors built two corpora based on journalistic texts. The first corpus (called AXIOLO) is produced by applying a set of manually extracted patterns compiled from 884 articles of the French newspaper *Le Monde* (on the topic "Portrait"). The second corpus (called ART) was generated from the same newspaper but on 3,539 articles talking about "Arts".

Results on both corpora have been reported. Results on the ARTS corpus show the interest of the selection tools in removing noisy patterns and selecting relevant ones.

[18] Since our work is also based on sequential data mining techniques and sequential patterns. More detail about such patterns, their properties and the methods used to extract them are thoroughly detailed in the next chapter.

In addition, based on this work the authors were able to discover new linguistic patterns not already reported in the set of the manually extracted patterns in order to identify qualifying appositive phrases. For instance, they discovered the pattern:

- $< (ADJ)(pour)(DET)(NOUN) >$ which matches phrases such as: "`célèbre pour son monastère`" or "`baroque pour une histoire d'amour`" alongside with also some other variations or extensions:
- $< (ADV)(ADJ)(pour) >$ which matches phrases such as: "`tres célèbre pour`"
- $< (ADJ)(pour)(VER) >$ which matches phrases such as "`indispensable pour assurer`"

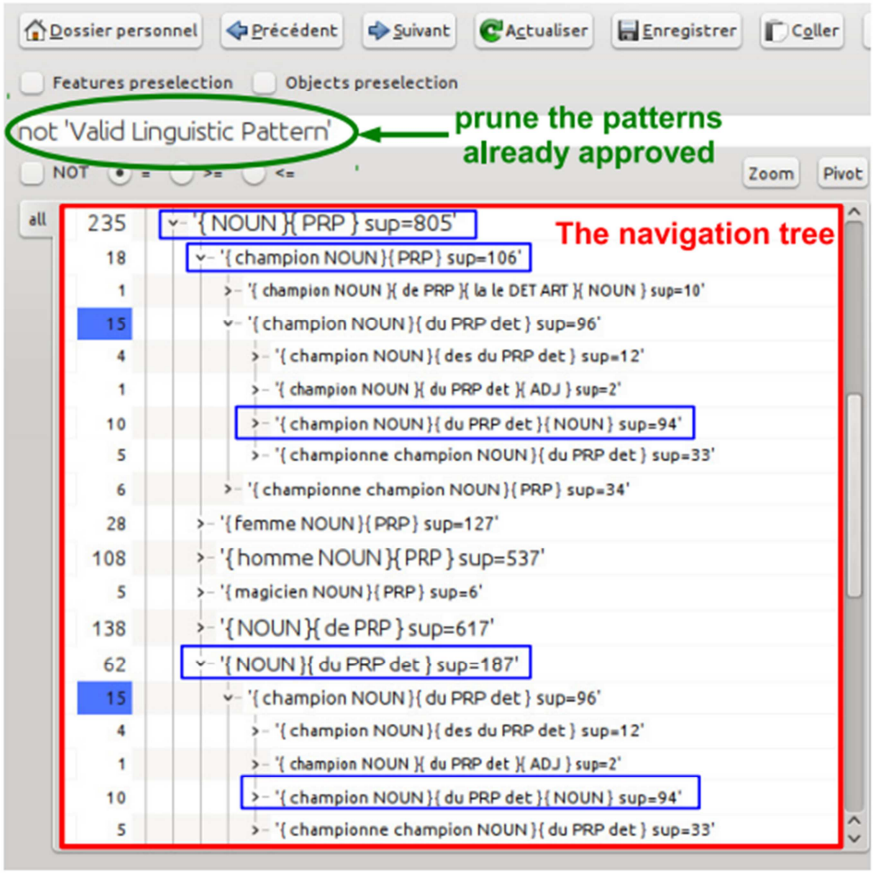

Figure 13. Example of pattern exploration and discovering with the ART corpus (Béchet et al. 2012)

Chapter 3. Considered Approach and Proposed Methods for the Extraction of Stylistic Patterns

As we have seen in the previous chapter, two main methodological approaches have emerged in the large field of computational stylistics both trying to bridge the gap between the statistical methods and techniques in the one hand, and the notion of writing style on the other hand: the hermeneutic approach and the classification approach.

As we have explained before, hermeneutic approaches are best suited for dealing with the notion of style and for extracting stylistic features capable of describing the stylistic aspects of a particular type of text without losing the very important point of interpretability, especially if we are looking for studying the style of a particular author in a relatively high level.

So, in our thesis we have been working on an approach to the computational stylistic study of French classic literature texts based on a hermeneutic point of view, where discovering interesting linguistic patterns is done without any prior knowledge or explicit a priori classification.

As explained before, in our work we focus on the development and the extraction of complex yet computationally feasible stylistic features that are linguistically motivated. We claim that the computational stylistic methods need to be grounded in the hermeneutic unsupervised paradigm rather than on the classification-based one. Following this line, we propose a knowledge discovery process for stylistic characterization with an emphasis on the syntactic dimension of style by extracting meaningful morpho-patterns from a given text.

Our aim is to assist linguists and literary researchers in studying the syntactic style and in extracting meaningful linguistic patterns from the text they are interested in. It is meant to support stylistic textual analysis especially from a syntactic perspective by:

1. Verifying the degree of importance of each extracted linguistic pattern (syntagmatic segments with gaps as we will see).
2. Automatically inducing a list of linguistic features that are significant, representative for an author's work.

In this chapter, we start by presenting in Section 3.1 overview of the proposed knowledge discovery process which can be considered as the core part of our thesis contribution. Before going into the details of the proposed knowledge discovery process, we give a very brief introduction to knowledge discovery that may help the reader not familiar with such domain to go through the chapter more smoothly. Section 3.2 and Section 3.3 are dedicated to the description of the two main steps of the proposed knowledge discovery process, namely the morpho-syntactic pattern extraction step and the interestingness assessment step.

3.1. Description of the Proposed Knowledge Discovery Process

This section gives an overview of the proposed knowledge discovery process and describes briefly the different steps that constitute it. Later, each step will be thoroughly detailed respectively in two separate sections.

As a general framework, the present study is inspired by interesting works such as (Ganascia 2002, Quiniou et al. 2012) where different texts are automatically analyzed in order to extract the most representative patterns from each of them based on a data-driven approach.

Indeed, corpus-driven and data-driven approaches differentiate from corpus-based approaches in that they make very little presupposition as to what to look for (Tognini-Bonelli 2001). In the present study, we do not seek to explicitly search for overuse or underuse of given predefined linguistic structures (e.g. nominalization, relativization, passive voice, etc.) in different literary texts but we let the corpus disclose what kind of structures (if any) are characterizing in each given text. Then, we use the knowledge we have of the text or its writer to interpret and evaluate the results a posteriori. Whereas most corpus-driven stylistic work done so far focuses on the lexical dimension of style, in our case we work on a pre-processed corpus annotated with POS tags.

In our contribution, we shall work both with purely syntactic patterns, and then lexical segments, namely with the textual instances of the syntactic patterns extracted from the analyzed text after the identification of the most representative ones.

It is evident that syntactic and lexical patterns do not necessarily act on the same stylistics level, since they capture different aspects of the linguistic choices that the author makes to create its texts. It is reasonable to formulate the hypothesis that the extraction of lexical instances of the syntactic patterns makes the analysis more sensitive to differences both in the content part of the text and the grammar, which makes the literary analysis and interpretation more complete and broader to different aspects of the style.

As opposed to the traditional method of extracting knowledge from data that relies completely on manual analysis and interpretation such as classic stylistics studies, knowledge discovery is the field concerned with the development of automatic and computational techniques for turning data into knowledge. More formally, knowledge discovery is a "*nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" (Fayyad et al. 1996).

The rapid growth of data collections across a variety of domains has made knowledge discovery more needed than ever in order to extract useful information from such huge amount of data that cannot be manually analyzed or handled by humans. Actually, this is the case for many fields where knowledge discovery is shown to be very useful and effective such as marketing, manufacturing or telecommunications.

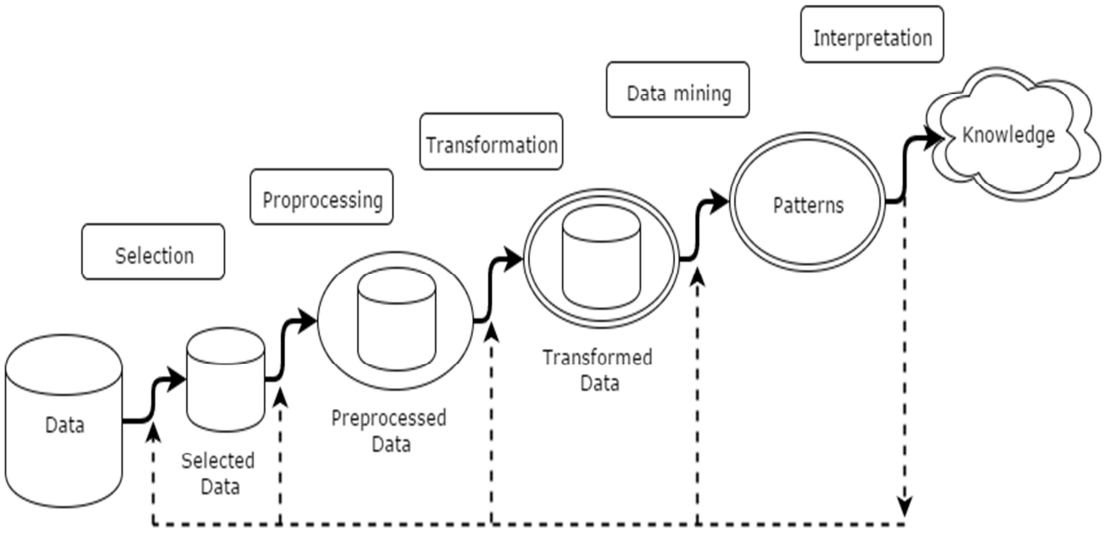

Figure 14. The typical pipeline of a knowledge discovery process

The knowledge discovery process is interactive in its nature, generally involving several steps some of which may require an intervention from the use. Figure 14 broadly outlines the usual knowledge discovery process prototype. For more theoretical details about its different steps, please refer to (Brachman & Anand 1996).

The proposed process can be qualified as data-driven hermeneutic in which texts are analyzed in order to automatically extract significant patterns that characterize the style of a certain text without any explicit a priori knowledge to be inserted as an entry to the process.

More formally, as illustrated in Figure 15 which represents it, the proposed process differs from the usual prototype since it is composed of two main automatic steps as explained below.

3.1.1. Morpho-Syntactic Pattern Extraction Step

In this step, each text is represented as a sequential database. Sequential patterns of a certain length are extracted using sequential pattern extraction algorithm (Viger et al. 2014). A morpho-syntactic pattern consists of a sequential syntagmatic segment (with possible gaps) present in the syntactic sequences[19]. This step consists in its turn of a pipeline, that is to say a sequence of different sub steps involving different processing tasks:

- Text cleaning
- Natural language processing (morpho-syntactic analysis)
- Sequential pattern mining

The details about this pipeline, the extraction procedure and the resulting morpho-syntactic patterns are thoroughly explained in Section 3.2 of this chapter.

3.1.2. The Interestingness Assessment Step

The pattern extraction step actually produces a huge number of patterns even from a small quantity of texts. However, the effective quantity of patterns that are actually of real interest is much smaller.

It is necessary to assess the relevancy of the extracted patterns in order to filter out the unimportant ones using some measures able to evaluate a pattern's actual worth depending on the application domain. Such measures are known as the interestingness measures.

As part of our thesis, we have proposed three interestingness measures based on different working hypothesis and implementing different ideas and interestingness properties. These measures are:

- Quantitative peculiarity-based measure
- Correspondence analysis-based measure
- Distribution peculiarity-based measure

Further details and explanation about these measures and their theoretical background are presented in Section 3.3 of the present chapter.

Clearly, this approach is more in line with the idea of an exploratory work, and gives some insights of on the opportunity of using such approaches to discover new facts about literary texts. It assume in the same time that the extracted patterns from the sequential data mining step should be source of stylistic knowledge if they are highly ranked by the interestingness measure step.
Therefore, our aim is to discover stylistic patterns that should ideally respect all the following properties:

[19] This is actually the definition in which we rely. However as explained in Subsection 2.6.2, there may be many other definitions for "syntactic pattern"

- To be capable of generalizing to new data (characterizing as many texts as possible)
- To be novel to the user or at least to confirm some already known knowledge
- To be useful to the user in achieving some related tasks. For our case, the ideal candidate task would be authorship and stylistic classification. That is to say, being able to identify the writings of a given author among different writings belonging to different authors (that would be a very interesting property for the patterns to have despite the fact that this is not our main purpose and expectation from such patterns)
- Finally, the patterns should be understandable in the sense that a user with a decent knowledge about the analyzed data would be able to interpret and understand those patterns

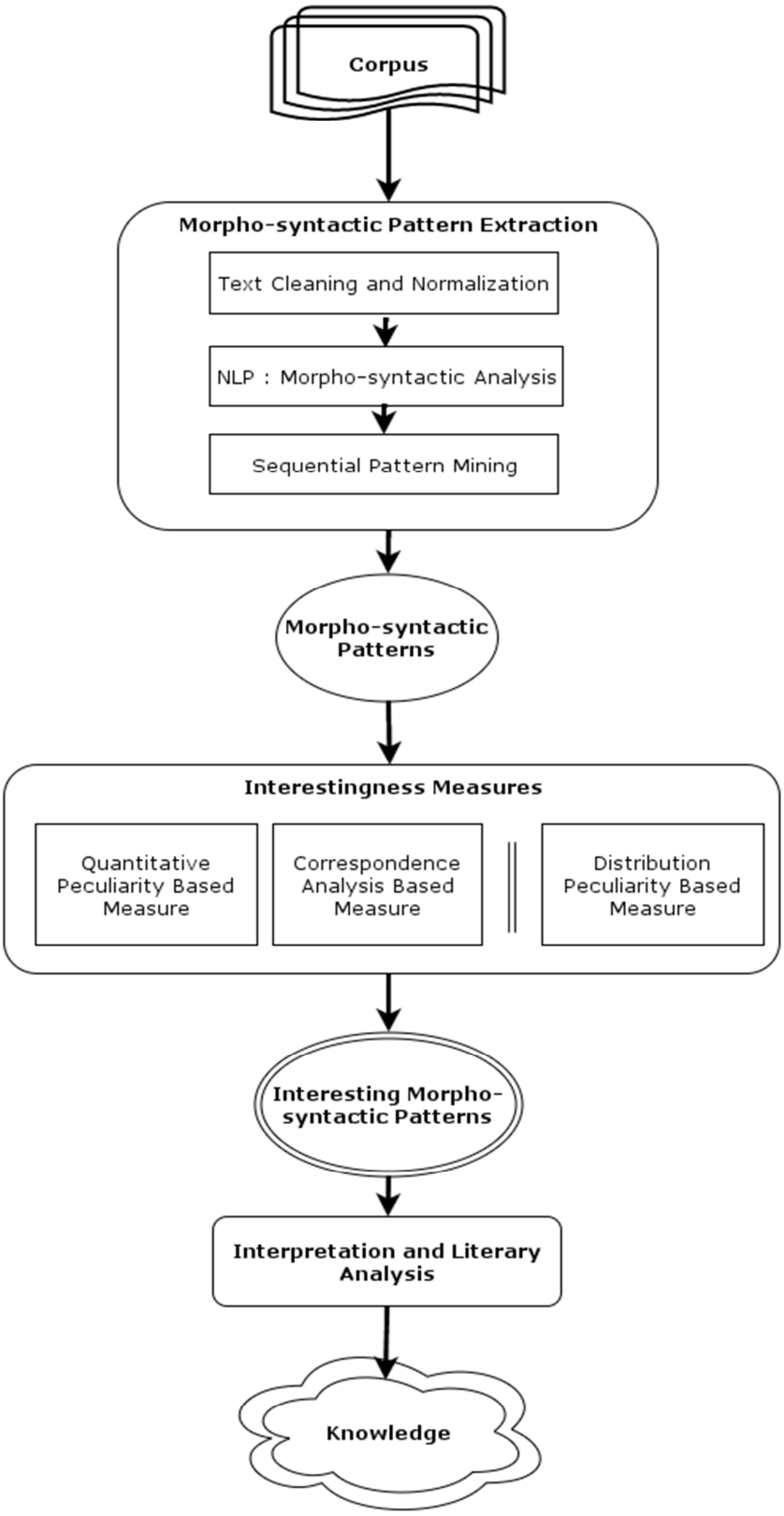

Figure 15. The proposed knowledge discovery process for the extraction of characterizing syntactic patterns

3.2. Extracting Morpho-Syntactic Patterns using Sequential Pattern Mining

Depending on its nature and on the information that it is carrying on, data can be possibly represented in different ways. For instance, it can be seen as graph, that is to say a certain number of elements (vertices) and edges representing some sort of connections or relationships between those elements. It can be also represented as a sequence of elementary information appearing in a certain order (imposed by time for example). Or in a more simple way, data could be basically represented as a set of raw quantitative and/or qualitative values without any kind of connections or order between them. To illustrate the variety and the complexity of data representation let's take the example of a corpus, namely a collection of textual documents.

Indeed, text documents are important sources of information, however many type of representations can be imagined for such collections. Each document can be viewed as a sequence of words and punctuation, and in which case the whole corpus will be seen as a set of sequences. A corpus can be viewed as a matrix as well, in which the rows represent documents and the columns represent some linguistic units' counts or frequencies. Actually, one can imagine many types of representations depending on the perspective in which the corpus is introduced and the application we want to do it with.

Foundational work in mathematics, probability theory, statistics and computer science provided an array of tools, techniques and processes used to explore data including textual data, which gave birth to the data mining domain and many other subdomains such as sequential data mining used to deal with sequentially represented data.

In this section, we describe the extraction of the morpho-syntactic patterns using sequential pattern mining which constitutes the first step of the knowledge discovery process presented previously. Subsection 3.2.1 sets up a theoretical background of sequential data mining in general with a special emphasis on sequential pattern mining in particular. Then, Subsection 3.2.2 makes the projection of sequential pattern mining problem to the computational stylistics domain by explaining the process in which the morpho-syntactic patterns are extracted. The last subsection (Subsection 3.2.3) describes lately in this chapter the properties of these patterns and explains the pertinence of the proposed interestingness methods.

3.2.1. Theoretical Background on Sequential Data Mining

3.2.1.1. Sequential Data Mining: Domain Introduction and Applications

Sequential data mining is a data mining subdomain introduced by (Agrawal et al. 1993) in order to deal with sequential data. It is concerned with finding interesting characteristics, rules and patterns in sequential databases. The problem of sequential data mining was first stated and defined for a commercial application. It was mainly motivated by the decision support problem faced by large retail companies (Stonebraker et al. 1993). Sales companies were already at that

time able to collect enormous quantities of sales data about theirs customers thanks to the democratisation of the code bare usage in the industry.

In such context, consider a database of customer transactions. Each transaction consists of the following characteristics: customer-id, transaction-time and the items involved in the transaction (the products purchased by this customer in this precise transaction). Such database is called sequence database. More precisely, each transaction is considered as an itemset (set of items) and each list of transactions having the same customer-id, ordered by transaction time, for the customer identified by that certain customer-id, is considered as a sequence. From such database and based on a decision-making point of view, many information and characteristics would be of interest for sales companies in order to study and understand the purchasing behavior of their customer, and by the way increase their sales' rates and benefits by developing a more customer-centred product strategies. A typical illustration of such interesting information mined from such database is that customers buying for instance item a than item b, have a big probability to buy the item c as well.

Lately, sequential data mining techniques were not only used to develop marketing and product strategies, but they were used to mine and extract meaningful information from other domains' databases such as telecommunication network alarm databases, intrusion detection (Hu & Panda 2004) and DNA sequences (Zaki 2003).

One of the domains with obvious special interest for us, and in which sequential data mining was also introduced, is computational stylistics. Actually, in an effort to develop more complex yet computationally feasible stylistic features that are more linguistically motivated, Hoover (2003) pointed out that exploiting the sequential information existing in the text could be a promising line of work. He proved that frequent word sequences and collocations can be used with high reliability for stylistic attribution. In another very interesting study, Quiniou et al. (2012) have shown the interest of sequential data mining methods for the stylistic analysis of large texts. They claimed that relevant and understandable patterns that may be characteristic of a specific type of text can be extracted using such techniques.

3.2.1.2. Sequential Data Mining: Problem Statement

In what follows, we will give a formal definition and problem statement of the sequential data mining problem. However, for the sake of clarity, we will limit our simplified definitions and annotations to those necessary to understand the experiments and the work done in this thesis. In fact, sequential data mining is a very large domain which involves many concepts and techniques beyond the scope of our interest.

Let's consider a set of literals called items, denoted by $I = \{i_1, \dots, i_m\}$. An itemset is a set of items. A sequence S is an ordered list of itemsets, denoted by $S = < s_1\, s_2 \dots s_n >$ where each s_j with $1 \leq j \leq n$ is an itemset. They are also called elements of the sequence. If each element in a sequence S consists of only one single item, S is called a single-item sequence. An item can occur multiple times in different elements of a sequence, but only one time at most in each element of a sequence.

For instance, the sequence $< (a\, b)\, (c)\, (d)\, (a) >$ is a sequence of four itemsets and $(a\, b)$ is an element of this sequence of two items a and b. The number of distinct instances of items in a sequence is called the sequence length. A sequence with length l is called a l −sequence

A sequence $S_1 =< a_1 a_2 \dots a_n >$ is included in another sequence $S_2 =< b_1 b_2 \dots b_m >$, i.e., S_1 is a subsequence of S_2 or S_2 is a supersequence of S_1, denoted by $S_1 \ll S_2$, if there exist integers

$1 \leq j_1 < \ldots < j_n \leq n \leq m$ such that $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots, a_n \subseteq b_{i_n}$.
For example, the sequence $< (a)(b\ c)(f) > \ll < (c)\ (a\ e)\ (a\ b\ c\ e)\ (e\ f) >$, since $(a) \subseteq (a\ e)$, $(b\ c) \subseteq (a\ b\ c\ e)$ and $(f) \subseteq (e\ f)$. However $< (a)\ (e)\ (f) >$ is not included in $< (a\ e\ f) >$ and vice versa.
A sequence database SDB is a set of tuples (id, S), where S is a sequence and id is its identifier. A tuple (id, S) is said to contain a sequence A if A is a subsequence of S. The support of a sequence S in a sequence database SDB is the number of tuples in the database containing S. The support of a sequence α in a sequence database SDB, denoted $supp(\alpha)$, is the number of tuples containing α in the database, defined as:

$$supp(\alpha) = \left|\left\{(id, S)\middle|\left((id, S) \in SDB\right) \wedge (\alpha \ll S)\right\}\right|$$

The relative support of a sequence can be defined as:

$$supp(\alpha) = \left|\left\{(id, S)\middle|\left((id, S) \in SDB\right) \wedge (\alpha \ll S)\right\}\right| / |SDB|$$

Let's consider the sequence database SDB given in Table 5 as a running example. The set of items included in the database are $\{ a\ b\ c\ d\ e\ f \}$.

Table 5. Sequance database SDB taken as runing example

Sequence-id	Sequence
S_1	< (a e) (b) (c) (d) >
S_2	< (a) (c) (d) (c e) >
S_3	< (a) (b) (c) (d) >
S_4	< (a) (c) (e) (f) >
S_5	< (c d) (c e) >

Many algorithms developed to extract and mine interesting characteristics can be extracted from such databases using sequential data mining:

Association Rule

Following the original definition by Agrawal et al. (1993), the problem of association rule mining is basically defined as: An association rule $R: X \Rightarrow Y$ is defined as a relationship between two itemsets X and Y such that $X \cap Y = \emptyset$. This rule can be interpreted as follow: if the items in X are contained in a certain sequence, the items in Y will be for sure contained in the same sequence as well.

For example, if we run an association rule mining algorithm on the SDB illustrated in Table 5, we will get as a result association rules such us: $(c\ e) \Rightarrow (d)$ with support equal to 2, which means that this rule is respected by two sequences in the SDB (i.e., there exist two sequences of the SDB where we find the itemsets $(c\ e)$ and (d) contained in the same sequence, in sequence S_2 and S_5 more precisely.

Sequential Rule

Quite similar to the association rule formulation with only one but very important difference, the problem of sequential rule mining is defined as: A sequential rule $R: X \Rightarrow Y$ is defined as a relationship between two itemsets X and Y such that $X \cap Y = \emptyset$. This rule can be interpreted as follow: if the items in X are contained in a certain sequence, the items in Y are also contained *afterward* in the same sequence.

For example, if we extract sequential rules from the SDB containing the five sequences presented in Table 5, we will get as a result sequential rules, such us $(a) \Rightarrow (c\ e)$ with support equal to 1, which means that this rule is respected by only one sequences in the SDB (sequence S_2), or $(a) \Rightarrow (c)$ with support equal to 4, (i.e., four exist three sequences of the SDB where we find the contained itemset (a), we find also the itemset and (c) contained afterward in the same sequences which are sequence S_1, S_2, $S_3$and S_4).

Sequential Pattern

In our context, the most important regularity that we are looking for in the SDM is the frequent sequential pattern. A sequential pattern is a sequence such that its support is greater or equal to a given user-predefined support threshold called $minsupp$. A frequent pattern is said to be maximal if it is not contained in any other frequent pattern.

In order to express the user interest about the most potentially interesting patterns, many other practical constraints can be incorporated such as the gap constraint. A gap represents the possibility to skip a certain number of itemsets between two itemsets of a sequence S. This gap is defined by two integers $mingap$ and $maxgap$ representing respectively the minimum and the maximum authorized size of that gap (the minimum and the maximum number of itemsets to be possibly skipped). A sequential pattern satisfying these two gap constraints is denoted by $P[mingap, maxgap]$.

For example, if we consider the $minsupp = 2$ in the running SDM of Table 5, the sequence $< (a)(b) >$ and $< (c)(d) >$ are considered to be a sequential patterns because they are contained in at least two sequences (S_1 and S_3) and three sequences (S_1, S_2 and S_3) respectively. Moreover, if we add the gap constraint into the extraction process, a different set of sequential patterns will shows up. For example considering $minspp = 2$ and $maxsupp = 2$ (exactly skipping two itemsets) the pattern $< (a)(d) >$ with $supp = 2$ will be the only $P\,[2,2]$ pattern in Table 5's SDM. ($< (a)(d) >$ is contained in both S_1 and S_3 while respecting the gap constraint).

3.2.1.3. Methods for Mining Sequential Patterns

Sequential pattern mining is a very interesting problem involving many challenges. First of all, one should know that a sequence database in practice includes a huge number of sequences, much larger sequences that those presented in the running example. A huge number of potential sequential patterns are included in such databases. Thus, mining algorithms should respect the following properties:

- Effectiveness: defined by being able extract the complete set of patterns without missing a single one
- Efficiency: being able to operate in a reasonable amount of time, which translates technically into the capacity to be scalable and to involve as less as possible of database scans

- Parametric use: being able to incorporate various kinds of execution parameter and user-specific constraints such as the length of the pattern and the minimum support

Many algorithms have been proposed to the efficient mining of sequential patterns or other frequent patterns in sequence databases. However, till the end of the 90s, almost all of the proposed methods for mining sequential patterns and other sequence-related patterns such as sequential rules were based on the *Apriori* algorithm (Agrawal et al. 1994). The basic idea exploited by the Apriori algorithm states the fact that any super-pattern of a non-frequent pattern cannot be frequent. Based on this heuristic, an Apriori-based algorithm method such as the famous GSP (Srikant & Agrawal 1996) adopts a "Candidate Generation - Candidate Pruning" approach. It performs in the following way:

First, perform a database scan to extract all of the frequent items which form the set of single item frequent sequences, namely the $1-$sequence seed set. Then, for each $k \geq 2$, use the $k-$sequence seed set to generate candidate $k+1-$sequences by pruning the $k-$sequence with respect to the Apriori property. Scan the database to find the support for each candidate sequence and keep only the candidates whose support in the database is no less than the minimum support. The remaining candidates constitute the $k+1-$sequence seed set.
The algorithm terminates either when no pattern's support above the minimum support is found in a pass, or no candidate sequence can be further generated. The Algorithms 1, 2 and 3 formalize this approach.

Algorithm 1. Apriori Algorithm

Apriori $(SDM, minsupp)$
  $L_1 \leftarrow$ {large $1-$sequence} ;
  for $(k = 2; L_{k-1} \neq \emptyset; k++)$ do
    $C_k \leftarrow$ New set of candidates generated from L_{k-1} ; // see Algorithm 2
    foreach sequence S in SDM
      Increment the support count of each candidate $c \in C_k$ if $c \ll S$;
    $L_k \leftarrow$ Candidates in C_k with support superior or equal to $minsupp$;
  return sequences in $\bigcup_k L_k$;
End Apriori

Algorithm 2. Apriori Candidate Generation

Apriori Candidate Generation (L_{k-1})
  $C_k \leftarrow \emptyset$;
  foreach sequence $p \in L_{k-1}$
    foreach sequence $q \in L_{k-1} \wedge q \neq p$
      if $(p.itemset_1 = q.itemset_1, \ldots, p.itemset_{k-2} = q.itemset_{k-2})$ **then**
        $C_k \leftarrow C_k \cup \{< p.itemset_1\ p.itemset_2, \ldots, p.itemset_{k\_1},\ q.itemset_{k\_1} >\}$;
  return C_k ;
End Apriori Candidate Generation

Algorithm 3. Apriori Candidate Pruning

Apriori Candidate Pruning (C_k, L_{k-1})
  foreach candidate $c \in C_k$
    foreach $(k-1)$-subsequence s of c
      if $(s \notin L_{k-1})$ **then**
        delete c from C_k;
  return C_k ;
End Apriori Candidate Pruning

If we go back to our running example presented in Table 5 and we iterate the algorithm over its tuples, the results representing the maximal sequential patterns will be such as illustrated in Table 6.

Table 6. Maximal sequential patterns resulting from the running example SDM

Sequence	Support
$< a\ b\ c\ d >$	2
$< a\ c\ e >$	2
$< d\ e >$	2

Even though the Apriori property and the generation-pruning mechanism reduce the sequential patterns' search space, Apriori-based algorithms have three main nontrivial drawbacks independently from their technical and algorithmic implementation (Han et al. 2000), briefly:

- The huge set of candidate sequences can be generated in a large sequence database which can dramatically slow dawn the mining process
- Many scans of the database in the mining process
- Performance and complexity issues when dealing with very long sequential pattern

Thus, starting from about the 2000 more original and technically efficient approach for mining sequential patterns have been proposed:

- Constraint-based sequential pattern mining: SPIRIT (Garofalakis et al. 1999)
- Pattern-growth methods: FreeSpan & PrefixSpan (Han et al. 2000, Pei et al. 2001)
- Vertical format-based method: SPADE (Zaki 2001)
- Mining closed sequential patterns: CloSpan (Yan et al. 2003)

3.2.2. Projection of the Sequential Pattern Mining to Computational Stylistics

As mentioned before, in our work we are interested in the extraction of characterizing stylistic patterns from certain texts, classic literary French texts more precisely.

Our aim is to extract morpho-syntactic patterns from that text. Sequential pattern mining is an appropriate technique to deal with such need. However, to do so we should also come up with an appropriate representation of the text in order to fit the input-output configuration of the mining process.

In our work, we have developed a processing pipeline (see Figure 16) to accomplish the extraction of morpho-syntactic patterns. The pipeline constitutes of a sequence of different steps involving different processing tasks: text cleaning, natural language processing, sequential pattern mining.

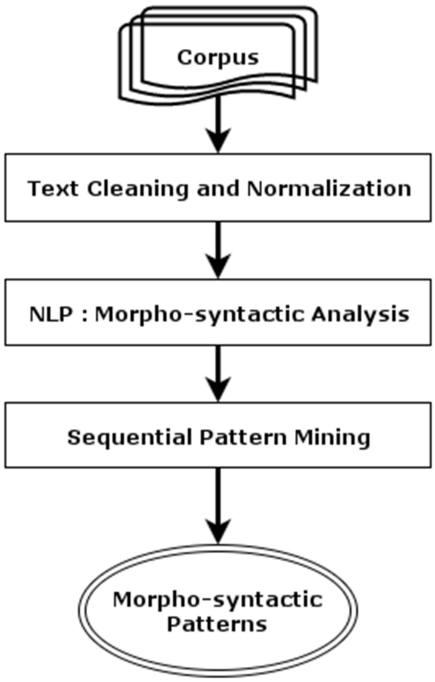

Figure 16. Different steps of the morpho-syntactic pattern extraction processing chain

3.2.2.1. Text Cleaning Step

Depending on its sources and formats, the text that we were required to handle in our work was more or less clean. Our main source was texts from Project Gutenberg[20] and Gallica[21] in two formats, either txt or xml format.

For xml-based sources, the task was to directly extract the main text written by the author of the work in question. For plain-text sources, the main task was to eliminate all the paratext, and all other material supplied by editors or publishers, which surround the main text. In both cases, we were also required to delete or normalize some special characters.

[20] https://www.gutenberg.org/

[21] http://gallica.bnf.fr/

3.2.2.2. Natural Language Processing Step

In our study, we consider a syntagmatic approach. We consider each text as a set of ordered sentences and each sentence as sequence of tokens, that is to say a sequence of syntactic constituents.

Basically, this step consists of transforming the cleaned text into a set of sequences of morpho-syntactic information destined to be analyzed afterward by the sequential pattern mining algorithm. This is done as follow:

1. The text is segmented into sentences based on strong punctuation. i.e., {. ? ! ...}
2. Each sentence is segmented into syntactic constituents called tokens
3. Each tokenized sentence is morpho-syntactically analyzed using a POS tagger, that is to say that for each token, in addition to the token, the tagger identifies its appropriate lemma and syntactic tag

This is done in such a way that each token will give birth to an itemset {tokens, POS tags, lemma} and consequently each sentence is seen as a sequence of itemsets. The whole text, which was already segmented into sentences, will produce a sequence database.

The choice of segmenting the text by sentences and not by other semantic unit is motivated by the fact that syntactic rules operate locally at the sentence level and do not necessarily extrapolate to larger and global coherent units such paragraphs (Akmajian et al. 2001).

To clearly illustrate the process done in the natural language processing step let's take the example of the following small textual segment extracted from Balzac's novel, *Eugenie Grandet*:

```
"La vie est une suite de combinaisons, et il faut les étudier, les
suivre, pour arriver à se maintenir toujours en bonne position. Charles
était un homme trop à la mode, il avait été trop constamment heureux par
ses parents, trop adulé par le monde pour avoir de grands sentiments."
```

So, the text is first segmented into two sentences, then each one of them is tokenized into syntactic constituents and part-of-speech tagged. As a result, the process produces the following two morpho-syntactic sequences for the two sentences respectively, each of which is constituted with a list of itemsets of the form {token, POS tags, lemma}:

Sequence 1:

```
< (La DET:ART le) (vie NOM vie) (est VER:pres  être) (une DET:ART un)
(suite NOM suite) (de PRP de) (combinaisons NOM combinaison) (, PUN ,)
(et KON et) (il PRO:PER il) (faut VER:pres falloir) (les PRO:PER le)
(étudier  VER:infi étudier) (, PUN ,) (les  PRO:PER La/Le) (suivre
VER:infi suivre) (, PUN ,) (pour PRP pour) (arriver VER:infi arriver) (à
PRP à) (se PRO:PER se) (maintenir  VER:infi maintenir) (toujours  ADV
toujours) (en PRP en) (bonne ADJ bon) (position  NOM position) (. SENT
.) >
```

Sequence 2:

```
< (Charles NAM Charles) (était VER:impf  être) (un DET:ART un) (homme
NOM homme) (trop ADV trop) (à PRP à) (la DET:ART le) (mode NOM mode) (,
PUN ,) (avait PRO:PER avoir) (été VER:impf être) (trop VER:pper trop)
```

```
(constamment ADV constamment) (heureux  ADJ heureux) (par PRP par) (ses
DET:POS son) (parents NOM parent) (, PUN ,) (trop  ADV trop) (adulé
VER:pper aduler) (par PRP par) (le DET:ART le) (monde NOM monde) (pour
PRP pour) (avoir VER:infi avoir) (de PRP de) (grands ADJ grand)
(sentiments NOM sentiments) (. SENT .) >
```

3.2.2.3. Sequential Pattern Mining Step

At this stage, each text is represented as a sequential database. Sequential patterns of a certain length along with their supports, which translate in our case to the number indicating how many distinct sentences contain the pattern, are extracted from this morpho-syntactic sequential database using a sequential pattern extraction algorithm (Viger et al. 2014). As we have briefly explained before, a syntactic pattern consists of a sequential syntagmatic segment (with possible gaps) present in the syntactic sequences. It can be considered as a kind of generalization of the notion of *skip-gram* used in the field of natural language processing. Ignoring the support information, here are some examples of syntactic patterns present in the sequence of the example above:

- `< (DET:ART) (NOM) (PRP) >`
- `< (pour PRP) (VER:infi)  (PRP) >`
- `< (PRP) (ADJ) (NOM) (. SENT) >`

Practically, the users of the processing pipeline have the possibility to specify different technical parameters:

- The minimum and the maximum length of the desired patterns
- The minimum and the maximum length of gaps allowed in-between the pattern's tokens
- The minimum relative support
- The minimum absolute support

In addition to this, some linguistic parameters could also be specified and taken into account in the extraction process:

- The possibility to restrict the extraction process only to the part-of-speech tags by omitting the tokens and the lemma information
- The possibility to restrict the analysis process to a reduced part-of-speech tag set. For instance in such tag set all the verb forms such as VER:infi (verb in the infinitive form) or VER:pres (verb in the present) are included in a higher tag containing them and representing the more general concept: VER (for verb) in this case.

3.2.3. Properties of the Extracted Morpho-Syntactic Patterns

3.2.3.1. Symbolic Property

From a symbolic point of view, the extracted patterns exhibited an interesting property. Actually, as noticed by Béchet et al. (2012) in their study of the linguistic patterns that describe quantitative and sentiment French sentences, such morpho-syntactic patterns are partially depending on specificity (see Figure 17). This is due to the fact that one POS-tag can cover many words, and in the same way, the lemmatization of different words can possibly results in one single lemma. Thus the textual instances matched by the pattern < `(pour PRP) (VER:infi) (PRP)`> can also be matched by the pattern < `(PRP) (VER:infi) (PRP)`> or just <`(PRP) (VER) (PRP)`> because these two latter patterns are more general patterns capable of covering the less general (more specific) former pattern.

This partial order property can represent a challenging issue for the analysis of the extracted patterns because it will produce some sort of information redundancy since one eventual stylistic textual segment can be covered by two or many morpho-syntactic pattern at the same time. For instance the textual instances "`pour arrive à`" or "`pour avoir de`" from the textual segment presented as an example above can be matched by all the three patterns cited above.

To deal with this issue, in our thesis we limit our analysis mostly to the general forms of the pattern, that is to say the pattern constituted only with the reduced part-of-speech tag set without including the token and the lemma information (in other words, the pattern that are on top of the lattice graph representing the partial order, < `(PRP) (VER) (PRP)`> in the lattice graph illustrated in Figure 17 for instance). The textual instances of the interesting patterns are extracted afterward from the text for the analysis and the interpretation needs.

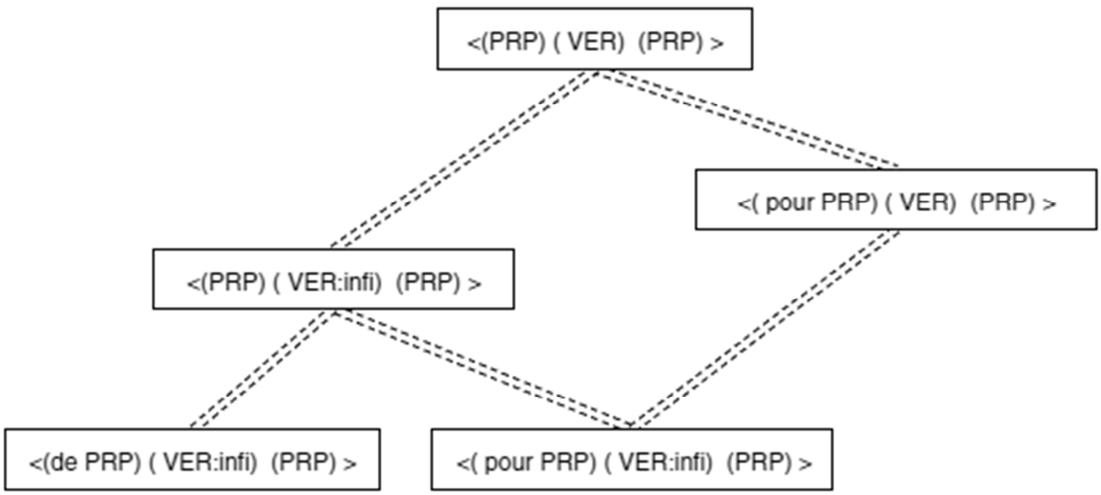

Figure 17. An example of partial order graph (lattice) involving five morpho-syntactic patterns

3.2.3.2. Quantitative Property

Sequential pattern mining is known to produce a large quantity of patterns even from relatively small samples of texts; however, it should be pointed out that the effective quantity of patterns that are actually of interest might be smaller. In fact, as noticed in a variety of domains, these extracted patterns exhibit a quantitative property known as the long tail distribution. Few patterns are very common (which translates into a high support), but most of them are quite uncommon (which translates into a relatively lower support). See Figure 18 as an illustration of this property.

Consequently, one cannot rely only and directly on the support of a pattern to determine its stylistics relevancy. This means that the importance and significance of a pattern should not be evaluated directly according to its counts in the texts. In fact, the long tailed distribution, as known in statistics, has tendency to decrease the significance of the low frequency events, even if they are implicitly relevant to the studied subject. This applies also in the context of linguistics (Montemurro 2001).

The property of the long tailed distribution on the one hand, and the statistical fluctuations on the analysis of patterns with low support values on the other hand, increase the weaknesses of methods using frequency-like-based measure (such as support) as main and direct element to evaluate the relevancy, and make them unsophisticated to discriminate the relevant linguistic forms in general.

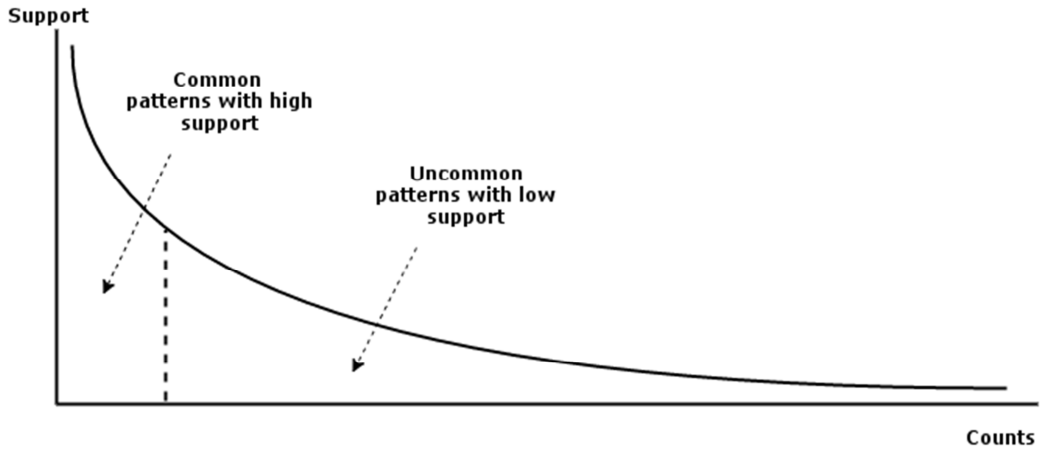

Figure 18. Illustration of the long tailed distribution property characterizing the extracted patterns

So, to deal with this issue and to avoid the effect of the statistical fluctuations on the analysis of patterns with low supports, in our thesis, we considered a minimum support threshold of 1%. That is to say that we focus only on patterns that are present in at least 1% of the sentences of the analyzed text. This threshold cut-off is claimed to be appropriate to have a kind of an a priori interestingness measure that filters the non-relevant patterns in the very beginning of the data mining step of the knowledge discovery process (Zhong et al. 2003).

However, some a posteriori interestingness measures should be applied on the remaining patterns in order to identify the most important and relevant ones. The three proposed interestingness measures are presented and discussed just in the next section.

3.3. Evaluating the Relevance of the Morpho-Syntactic Pattern using Interestingness Measures

As we have seen in the previous section, data mining in general and the sequential data mining in particular can generate huge number of patterns from data, hundreds and often thousands of patterns depending on the size of the sequences database. It is therefore necessary to be able to identify patterns that are of actual interest.

In this case, in order to have a useful knowledge discovery system able to enhance the knowledge of the system's user about the studied text, it is necessary to assess the relevancy of the extracted patterns. Basically, this can be done by determining the most relevant patterns from those that are not necessarily relevant. In other words, it is necessary to filter out those patterns using some measure of the patterns' actual worth depending on the application domain and the knowledge mined from this domain data. These measures are often known in the literature as the interestingness measures.

In this section, we introduce the proposed interestingness measures used to assess the relevancy of the extracted morpho-syntactic patterns. This constitutes a major part of our contribution. The section is organized as follows: Subsection 3.3.1 presents the theoretical aspect about interestingness measures. After that, Subsection 3.3.2 discusses the proposed three interestingness measures and gives details about their linguistics motivations and statistical formulations respectively.

3.3.1. Theoretical Aspects about Interestingness Measures

3.3.1.1. Interestingness Measures Categorization

As one can sense, the notion of interestingness is very wide and opaque. Hence there exists no consensus on widespread formal definition of interestingness agreed among the knowledge discovery community. In fact, the notion of interestingness is quite large and related to the domain context. For those reasons, measuring the interestingness of patterns produced by data mining techniques is not an achieved task and still an active research area in the field of knowledge discovery.

However, these interestingness measures can be generally divided into two categories: objective and subjective measures.

1. Objective interestingness measures are based on the idea that no knowledge about the studied domain is required. So, ideally objective measures are based only on the statistical properties of the discovered patterns

2. Subjective interestingness measures, as opposed to objective measures, take into consideration both the properties of the discovered patterns and the knowledge that are already known about the concerned domain. Thus, to conceive a useful subjective measure, access to the user's knowledge about the studied domain in general and more particularly about the mining task is required. Practically, this can be done in two different manners. Either by asking the user to intervene within the knowledge discovery process, or by explicitly representing the user's knowledge and injecting it as an input to the knowledge discovery process.

3.3.1.2. Properties of Interestingness Measures

Geng & Hamilton (2006) have considered nine criteria for determining whether a pattern is interesting or not. They propose to treat the interestingness as a broad concept that emphasizes: conciseness, coverage, reliability, peculiarity, diversity novelty, surprisingness, utility, and actionability.

These nine criteria that are more or less specific to the nature of the assessed patterns are used to measure whether or not this pattern is interesting. They can be described as follows:

- Conciseness: a pattern is concise if it contains relatively few attribute, while a set of patterns is concise if it contains relatively few patterns, which makes them easier to understand and remember
- Generality/Coverage: a pattern is general if it can generalize to as much as many instances to cover a relatively large subset of a dataset. This property can be captured by the support in the case of sequential data mining
- Reliability: a pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases
- Peculiarity: a pattern is peculiar if it exhibited a different behavior from other discovered patterns according to some distance measure. Peculiar patterns have more chance to be unknown to the user, hence interesting
- Novelty: a pattern is novel if it is unknown to the user
- Diversity: a pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if its constituent patterns differ significantly from each other
- Surprisingness: a pattern is surprising either if it contradicts the user existing knowledge or was already known but not expected to be extracted
- Utility: a pattern is of utility if it can contribute to reaching some goals concerning the knowledge that can be extracted from a dataset
- Actionability: a pattern is actionable in some domain if it enables decision making about future actions in this domain

Clearly, some of these nine criteria need some sort of access to the user's domain knowledge about the mined data, which makes the measures implementing them to be considered more or less as subjective (criteria such as novelty and surprisingness). By contrast, such criteria as conciseness,

generality, reliability and peculiarity, which depend mainly on the statistical and symbolic properties of the extracted patterns themselves, tend to be more objective.

3.3.1.3. The Roles of Interestingness Measures

During the knowledge discovery process, interestingness measures can be used for two main roles (see Figure 19):

Firstly, they can be used to prune uninteresting patterns from the very beginning of the process during the data mining step so as to reduce the amount of the extracted patterns and thus improve the efficiency.

Typically, the generality/coverage interestingness criterion, which translates in most cases (as in our's) into a support's minimum threshold, could be used to avoid extracting patterns with low supports. Similarly, some other criteria could also be used to prune patterns in the same manner.

Secondly, measures can be used during post-processing to select interesting patterns. Two different manners for determining whether a pattern is interesting can be imagined:

1. On the one hand, one manner would be to directly classify each extracted pattern as being either interesting or uninteresting. Then, only the set of interesting patterns will be considered for a further analysis or usage
2. On the other hand, we can rank the patterns with respect to some interestingness value. So, no formal decision about the interestingness is made, only an order suggesting that that one pattern is more interesting than others. Actually, this is the preferred manner in our approach, so we will be relying on such mechanism in order to automatically evaluate the importance of the extracted patterns. Simply put, the higher ranked the pattern is, the more important it is.

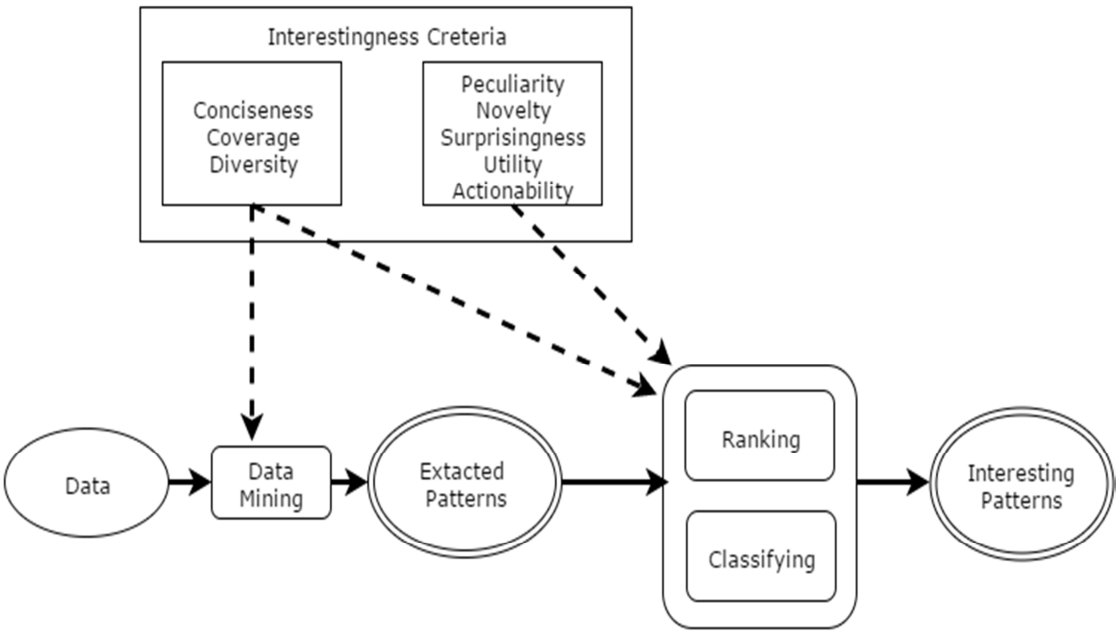

Figure 19. The roles of interestingness measures (Geng & Hamilton 2006)

3.3.2. Proposed Interestingness Assessment Measures

As part of our thesis, we propose three interestingness measures based on different working hypothesis, and implementing different ideas and interestingness properties. However, they are all considered to be objective measures since there is no domain knowledge explicitly represented in the formalization of the measures. The main interestingness criterion implemented by these measures is *peculiarity*. However, each one of them takes a different perspective to assess this peculiarity.

What is interesting about the peculiarity is that it is by definition the most likely criterion to be captured with some statistical procedures. More concretely, what these measures do is that they evaluate automatically the behavior of each extracted pattern (in a specific perspective) with respect to the other patterns' behavior. In other words, the extracted patterns are ranked with respect to some peculiarity-assessment measure. The more peculiar (highly ranked) the pattern is, the more likely it is to be worth further investigated by the user, hence interesting.

The goal of this part is to present these practical measures for extracting relevant syntactic patterns from texts for stylistic analysis purpose. They are motivated by both statistical and linguistic considerations. Since they do not rely on the raw support of the syntactic patterns in texts, these measures can work reasonably well with both large and small text samples and allow the extraction of significant syntactic patterns from a stylistic point of view.

The three interestingness measures presented in the next three subsections respectively are:

1. Quantitative peculiarity-based measure
2. Correspondence analysis-based measure
3. Distribution peculiarity-based measure

These three measures cluster into two different categories: extrinsic and intrinsic measures as shown in Figure 20.

The first two ones (quantitative peculiarity and correspondence analysis-based measures) are both extrinsic measures in the sense that they are implementing a comparative methodology. That is to say that the interestingness of a text or of some of its elements (syntactic patterns in our case) is measured by comparison to other texts. Basically a text needs to be compared to other texts or to some comparative corpus (if available) in order to extract what is special about it.

As opposite to them, the third measure (distribution peculiarity-measure) is an intrinsic measure in the sense that it does not implement a comparative methodology. The interestingness of syntactic patterns extracted from some text is not measured by comparison to other texts, but strictly based on some inter-textual properties of those patterns. So, no other texts or comparative corpus are needed.

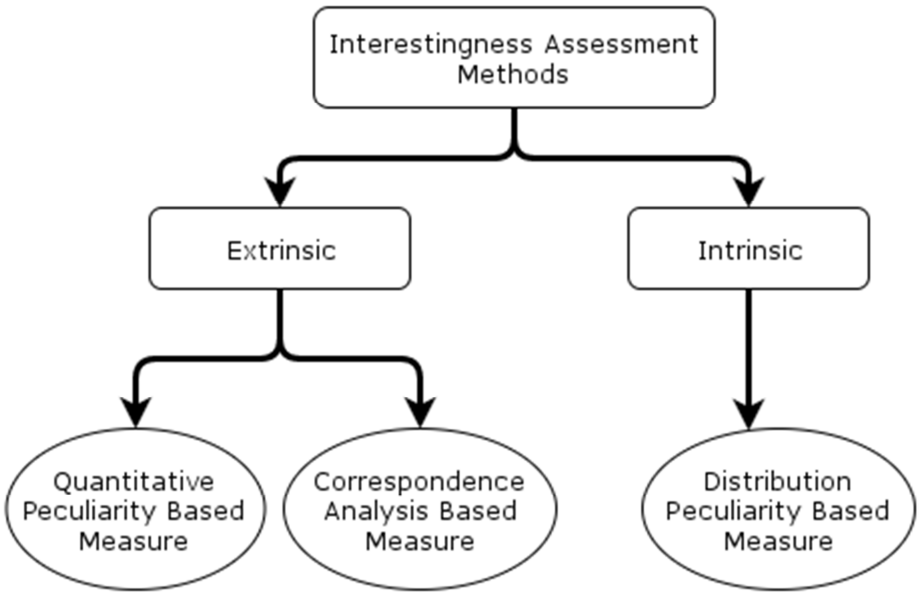

Figure 20. The three interestingness measures proposed and presented in our work

3.3.2.1. Quantitative Peculiarity-based Measure

In this subsection, we present a quantitative peculiarity-based interestingness measure that evaluates the overrepresentation (over-employment in terms of support with respect to a comparative corpus) of extracted syntactic patterns resulting from the sequential pattern mining step. Thus, each syntactic pattern will be assigned an interestingness value indicating its importance and its relevance for the characterization of text's syntactic style.

Hypothesis for Evaluating the Patterns Relevance

Our hypothesis to evaluate the relevance of a syntactic pattern is based on the fact that the most relevant ones should significantly reflect the stylistic choice of the author and should thus be characterized by significant peculiar quantitative behavior. This peculiar behavior translates into a support's over-representation in author's texts. However, to capture this overrepresentation one cannot refer only to the absolute frequency of occurrence (support). Indeed, more frequent use of a syntactic pattern by an author (patterns with a relatively high support) does not necessarily indicate a stylistic choice since it can be very well a property imposed by the grammar of the language or the text's genre.

Mathematical Formulation and Assessment of the Relevance

Thus, to assess the over-representation of a pattern, we use an empirical approach based on the comparison of the support of a syntactic pattern in a text to that found in a comparative (norm) corpus. A ratio α between these two quantities is calculated as follow:

$$\alpha = \frac{\text{Pattern's support in the text}}{\text{Pattern's support in the norm corpus}}$$

In our experiments we found empirically that the distribution of the ratio α exhibits a Gaussian behavior. Indeed, the values of the α ratio are normally distributed around a central value (see Figure 21). This is due to the fact that the frequency of occurrence of a syntactic pattern in a text is highly correlated with the frequency of occurrence in the norm corpus, with a few exceptional special cases or outliers (see Figure 22). These outliers represent the patterns of special interest for our study because they represent a certain linguistic deviation that is specific to the author's style compared to what one would expect to see in the norm corpus.

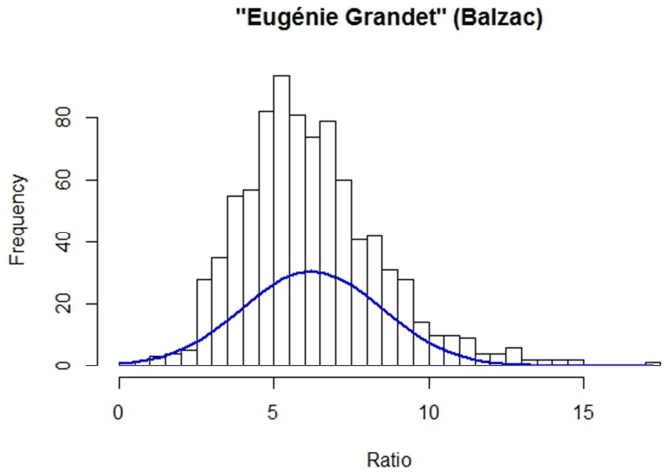

Figure 21. Illustration of the Gaussian behavior of the ratio α in Balzac's *Eugénie Grandet* novel

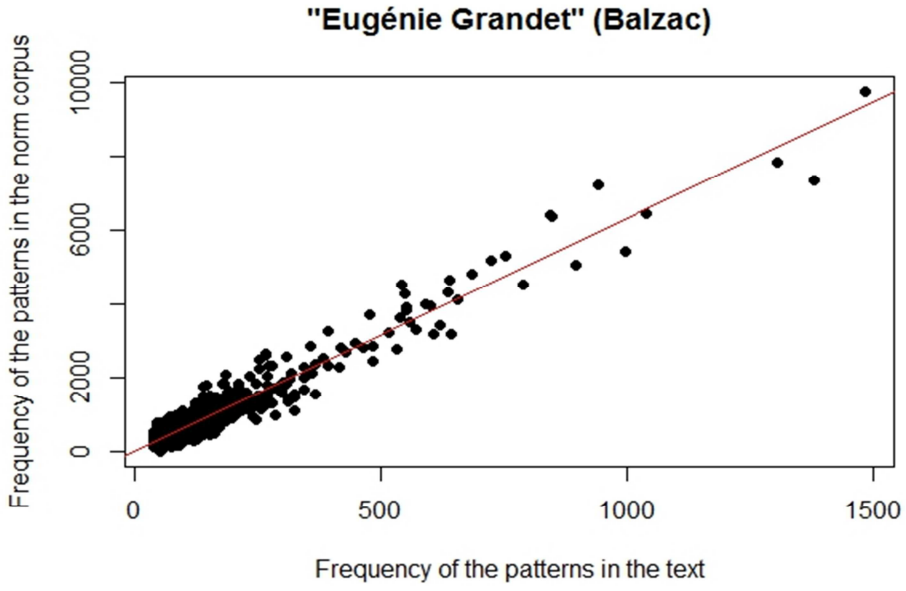

Figure 22. The supports of syntactic patterns in a text with respect to their supports in the norm corpus. Each point in the graph represents a syntactic pattern. The plotted line represent the linear regression line capturing the expect quantitative behavior

The configuration described above allows us to use an outlier detection method based on Gaussian distribution and Z-score to identify such special patterns (Chandola et al. 2009). The over-representation of a pattern in this case will result in a greater positive aberrant behavior compared to other patterns. The most over-represented patterns will be those associated with highest values of standard z-score Z. The z-score values are calculated as follows:

$$Z_i = \frac{\alpha_i - \hat{\alpha}}{S}$$

Where α_i and Z_i are respectively the ratio α and the z-score corresponding to the i-th syntactic pattern. $\hat{\alpha}$ and S are respectively the mean and standard deviation of the ratio α .

3.3.2.2. Correspondence Analysis-based Measure

In this subsection, we present a different extrinsic method for the exploration of interesting syntactic patterns. This method is based on the evaluation of the contribution of the syntactic patterns when using correspondence analysis projection on the studied texts.

Basically, to do so, each text in the analyzed corpus is represented as a vector of supports of the syntactic patterns extracted from it after the sequential pattern mining step. Correspondence analysis is a dimensionality reduction technique (Benzécri 1977) that is a well-known and often used in digital humanities and textual analysis (Lebart et al. 1998). The main advantage in using correspondence analysis with respect to the previous method is that the complete results of the analysis are available in a series data structure rather than a single real value. It allows for a selective printing of a subset of patterns on the plot; moreover the proximity of a pattern to any of the texts can be easily calculated by some distance function such as Euclidean distance, thus allowing for the automatic filtering of patterns more strongly associated with one text than to the others.

Hypothesis for Evaluating the Patterns Relevance

The most important result is the contribution of each pattern on the two axes; it is defined as the actual contribution of that pattern to the overall displacement of the position of texts in the resulting plot. If a pattern is strongly characterizing of a text with respect to the others, it will contribute greatly to the displacement of the text in the bi-dimensional space. Thus, the average contribution of such a pattern on the two axes of this pattern will be higher than the one of other patterns that has more or less the same frequencies in all texts. Subsequently, the main idea here is to use the contribution as an interestingness measure to rank patterns.

Assessment of the Relevance

To understand the positioning of patterns and texts in a bi-plot (see Figure 23 as an illustration), the metaphor of a magnetic field can be used. The majority of patterns are concentrated in the center, because they are equally attracted (represented) in all texts. On the other hand some patterns are strongly attracted by just one text and are repulsed by the others, positioning themselves at the extremity. Others are equally attracted by two texts only, positioning themselves somewhat in between. Moreover the force of attraction is not the same. Some patterns seem to be stronger in pulling a text towards them. For instance, in Figure 23, Balzac's and Zola's points (representing their texts) are less central. This can be interpreted in the sense that such

texts have stronger characterizing features than the other two. Ideally, it would be interesting to be able to select for further analysis only the pattern that are actually more contributive for the displacement of the texts over the two axis. The correspondence analysis statistical procedure provides us with a contributive value for each pattern in the plot. By combining contribution values and proximity, it is possible to select, among the patterns with high contribution, those that are nearer to one text than to the other three. This is calculated by measuring some distance function between the position of each text and the feature (the pattern), and choosing the nearest text. Many distance function can be used in practice. In our experiment we decided to use the Euclidian distance which seems to be more appropriate for our need.

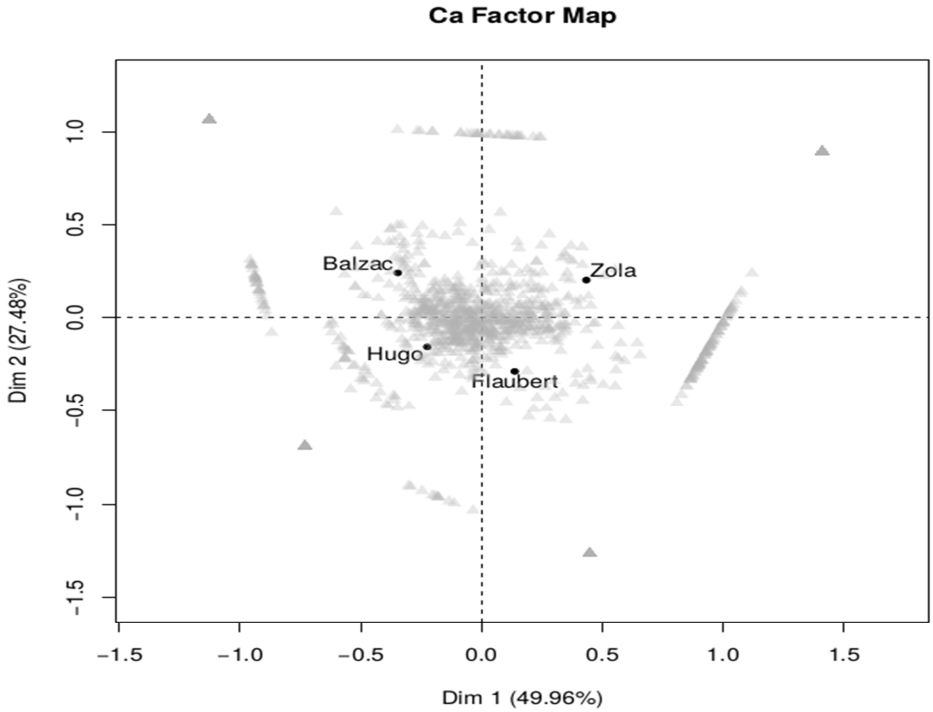

Figure 23. Plot with partially out-shadowed patterns. In the image, patterns are unlabeled and represented in grey, while texts are represented in black and labeled with their author's names

3.3.2.3. Distribution-Peculiarity based Measure

In this subsection, we present the intrinsic method for assessing the interestingness of the morpho-syntactic patterns. By opposite to the first two intrinsic methods, to assess the relevancy of a given morpho-syntactic pattern, this interestingness measure is based on the position in which that pattern appears in the text (The distribution of the pattern in the text), rather than its support.

Hypothesis for Evaluating the Patterns Relevance

Our hypothesis is based on the idea that the occurrences' positions of the most characterizing linguistic patterns should be controlled by the author's purpose, while the irrelevant linguistic patterns are distributed randomly in the text. The assumption made in this measure is that the higher the importance of a linguistic pattern is, the more its occurrences cluster together detaching them from a random distribution. By this methodology, we search for patterns whose frequency is much higher in single portions of texts than in others, thus making each of them the locally most prominent pattern.

The clustering phenomenon can be visualized in Figure 24 where we have plotted the absolute positions in the text of two different syntactic patterns. In this code bar representation, the left edge of the bar represents the beginning of the text; the right edge represents its end. A very thin vertical line is drawn at the position of each occurrence. One can clearly notice that despite the two patterns having the same support (counts of sentences where they appear); they significantly behave differently in terms of their distribution of occurrences' positions. This property gives them a different linguistic relevancy value.

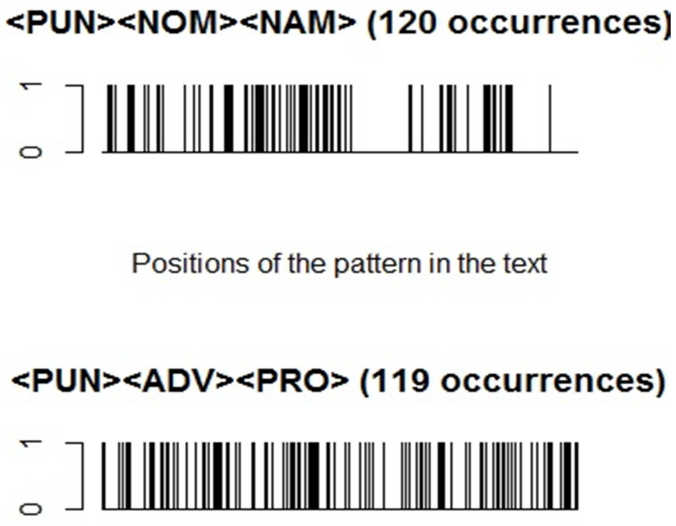

Figure 24. Positions of occurrences in the text, counted by sentences from the first to the last one, of two different patterns with approximately same support, but with different distribution of positions.

Mathematical Formulation and Assessment of the Relevance

The positions of the pattern's occurrences, with support equal to n in the text, are denoted by $P = (p_1, p_2, p_3 \ldots p_n)$ where p_i is the rank of the sentence in which the pattern appears for the i-th time with $i \in (1, .., n)$. The set of distances between two successive occurrences of a pattern can be denoted by $D = (d_1, d_2, d_3, \ldots, d_{n-1})$ where $d_i = p_{i+1} - p_i$.

Given this configuration, in order to quantify the degree of importance of each linguistic pattern and thus evaluate its relevance based on its clustering behavior, we use the parameter σ which is defined as the standard deviation of the distribution of the set of distances D normalised by the expected value of d such as:

$$\sigma = \frac{\sqrt{E(D^2) - E(D)^2}}{E(D)}$$

For the case where the distribution of the pattern position set P is completely random, one should expect the corresponding distances set D to follow a Geometric distribution and thus to have a parameter $\sigma = 1$, and the larger is the clustering the bigger is σ (Ortuno et al. 2002). However, patterns with different random distributions of D (different random clustering settings in a text) would have a significant difference in their corresponding parameter σ. To avoid this, we normalize the pattern parameter σ with the Geometric parameter σ_{geo} where:

$$\sigma_{geo} = \sqrt{1 - (1/E(d))}$$

Such as (σ_{norm} is the normalized σ):

$$\sigma_{\text{norm}} = \frac{\sigma}{\sigma_{geo}}$$

Such method was already successfully used in physics to quantify energy level of disordered system and in information retrieval to extract key words and keys phrase from informative texts (Carpena et al. 2009).

Chapter 4. Experimental Evaluation and Results

The evaluation is a crucial step of any research process including the exploratory one. However, such paradigm of research is very particular, and differs considerably from other paradigms in many aspects including the analysis, the formulation and the evaluation. In fact, evaluating results coming from unsupervised exploratory researches is a very hard task unless well formalized a priori knowledge, to be taken as an evaluation material, is available. Unfortunately, this is not the case in most applications. This somehow makes sense, since relying on exploratory approaches most probably induces that no well-established knowledge is already available or formalized about the studied domain.

In fact, this is also the case for our work. As we have clarified in Chapter 3 discussing the approach in our work, we are completely relying on an unsupervised knowledge discovery process that does not necessitate any a priori explicit knowledge to be taken into account.

Actually, this is in line with methodologies used in literary studies (and cultural sciences) where research activities are mostly dominated by intuition and non-formalized knowledge (knowledge that cannot be directly exploited by computers). On the other side, such methodology prompts the question of the evaluation and how it should be handled.

Well, in our case, the ideal situation to have is to acquire what is called the *ground truth* (knowledge provided by direct observation as opposed to knowledge provided by automatic ways). In other words, this means that one should have a set of manually extracted morpho-syntactic

patterns describing the style of a corpus or an author. The extraction should be done by a domain expert in order to have credibility, that is to say, a literary or a linguistic scholar specialized in the corpus or the writings the author in question.

Then, one should compare the results produced by the automatic extraction systems with the ground truth. Some evaluation measures are computed afterward to assess how much the automatic extracted results match the ground truth. This gives an estimate of the automatic extraction system's performance.

Unfortunately, such ground truth knowledge does not exist in our case, and more sadly it is quite very hard, even impossible, to acquire for many reasons.

First of all and as we stated before, literary research activities (including stylistics) are intuition-based. The literary researcher in most cases can sense the stylistic aspects of a given text, but he/she cannot formalize it in such a way that can be taken into account in an automatic evaluation task. Actually, this makes sense and this is where the strength of computational stylistics comes from, since it aims to find patterns linked to styles which are not demonstrable without computational methods (Craig 2004).

Secondly, to be as objective as possible, one should not rely on a ground truth annotation given by one expert, and not for only one text. The ground truth annotation should be as large as possible (many experts annotating many texts) in order to accurately evaluate the generalization performance of the proposed automatic system. This assumes that more than one literary expert, having decent knowledge in the text under investigation, should be available to participate in the ground truth annotation operation. This assumption is very strong and so hard to satisfy; it just does not apply in practice. Indeed, computational stylistics is a very scattered domain containing several subtasks that are not that well formalized. Thus, there exists at our best knowledge no evaluation benchmark available (not necessarily extracted by human experts, but at least validated by them) for the stylistic characterization with which one could objectively compare the quality of the produced results.

Finally, as a generic property to all the annotation operations, such tasks are known to be hard to accomplish and they are very time-effort-consuming.

For all those reasons, in our work, we will rely instead on a different evaluation protocol. Indeed, the evaluation considered in our work consists of two parts:

Qualitative evaluation

In the first part, we conduct a qualitative evaluation of the patterns extracted using the proposed knowledge discovery process. The experimental results are given by applying respectively the three proposed interestingness measures on the morpho-syntactic patterns (constituted of 3 to 5 itemsets, 3-5 grams) extracted from a corpus meant voluntarily to be small.

For each interesting-ness measure and for each text in the analyzed corpus, we selected the top 10 most relevant patterns. Then, linguistic and stylistic interpretations are made on each selected pattern with the help of linguistic and literary researchers from the Labex OBVIL[22]. In that matter, The OBVIL's literary researchers were provided with a data collection containing the most 10 relevant patterns identified by the three proposed measures for each text respectively.

[22] http://obvil.paris-sorbonne.fr/

The data collection contains also the textual instances of each pattern in the corresponding text. Based on their literary and stylistic knowledge about the studied texts on the one hand and this data collection (relevant patterns and their textual instances) on the other hand, the researchers have made stylistics and literary interpretation assessing the quality of the extracted patterns and their corresponding measures. From our preliminary experiments, we find out that the resulting morpho-syntactic patterns (word form combined with lemma and part-of-speech tag) are much numerous and difficult to interpret. Indeed, it was very hard for literary researchers with whom we have worked to make sense from such data. To deal with such issue in this part of the evaluation, we limit our analysis to the most general forms of the pattern, that is to say the pattern constituted only with a reduced part-of-speech tag set without including the token and the lemma information. This evaluation can be seen as a qualitative analysis of the extracted patterns.

Quantitative evaluation

In that part, we will quantitatively evaluate, in a larger corpus than the previous one, whether the extracted patterns are suited to differentiate the writings of given author from another one using a clustering algorithm. In that matter, among other textometric features used for a comparative purpose, the extracted patterns are used as features to describe the text. Our aim is to have a quantitative assessment of the discriminant power of these patterns (extracted as characterizing the style) given the fact that the stylistic choices of an author are the elements that can allow us to distinguish his writings from others.

4.1. Qualitative Evaluation

In this section we present the qualitative analysis done on the extracted patterns using the three interestingness measures of the proposed knowledge discovery process. The experiment results are given by applying them on the set of extracted morpho-syntactic patterns.

First, in this section we start by presenting the corpus used for the qualitative evaluation in Subsection 4.1.1. Then, we proceed to the presentation of the qualitative results for each one of the three proposed interestingness measures in Subsections 4.1.2, 4.1.3 and 4.1.4 respectively (morpho-syntactic patterns extracted as relevant by the corresponding measures and their literary and linguistic interpretation).

4.1.1. Analyzed Corpus and Experimental Settings

In our study, we used four novels, belonging to the same genre and the same literary time span, written by four famous classic French authors: Balzac's *Eugenie Grandet*, Flaubert's *Madame Bovary*, Hugo's *Notre Dame de Paris* and Zola's *Le ventre de Paris* as illustrated in Table 7. This choice is motivated by our particular interest in studying the style of the classical French literature of the 19th century. Moreover, we believe that such choice helps us to focus more on the individual style of each author by limiting the effect of the genre and its functional impact.

Table 7. The analyzed corpus for the qualitative evaluation

Author	Work	# of word
Balzac, Honoré de	*Eugénie Grandet*	62849
Flaubert, Gustave	*Madame Bovary*	111109
Hugo, Victor	*Notre dame de paris*	168624
Zola, Émile	*Le ventre de Paris*	110558

4.1.2. Quantitative Peculiarity Results and Discussion

In this subsection, we present some examples of relevant syntactic patterns extracted from the analyzed corpus. At the time of the analysis of the syntactic patterns, each text written by one of the four authors is contrasted with texts written by the three other authors. That is to say that these three texts will be considered as the norm corpus against which we will evaluate the hypothesis of the overrepresentation of syntactic patterns in the fourth remaining text as explainned later in this subsection.

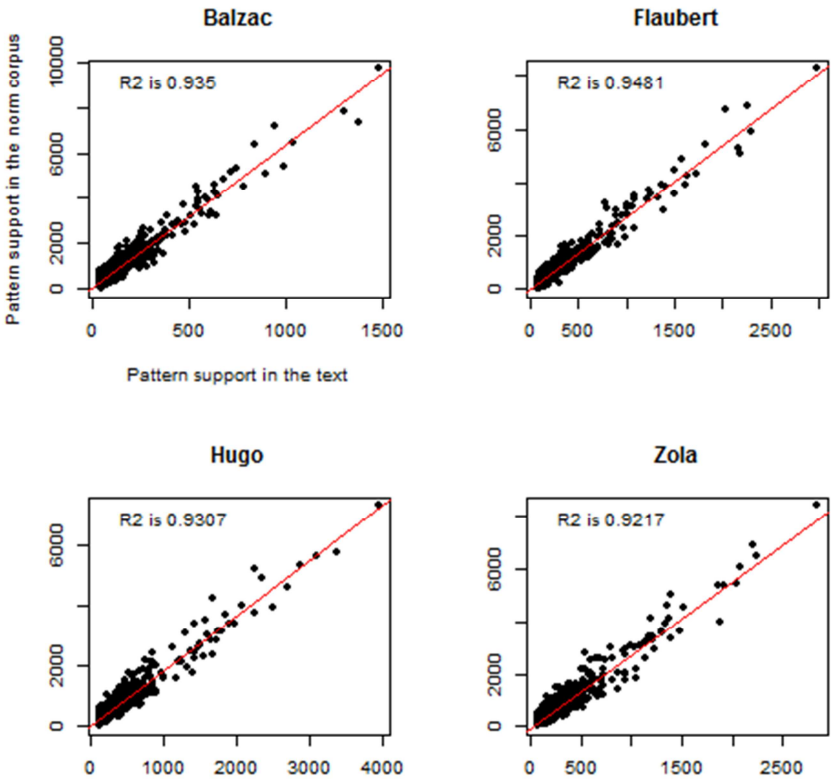

Figure 25. Contrast of the syntactic patterns' support in each text with respect to their frequencies in the whole corpus

Figure 25 illustrates the plot resulting from putting in contrast the support of the patterns in each of the four studied novels (x axes), labeled in the plot with the names of their authors, with respect to their support in the norm corpus (y axes).

Each point in the graph represents a syntactic pattern. The plotted lines represent the linear regression lines capturing the expected behavior of α *ratio*[23].

As we can notice from this plot, the four novels exhibit almost the same behavior and patterns' configuration. Actually this translates statistically into an R^2 score almost similar for the four novels (statistical measure, presented in the top left of each individual plot, of how close the data are to the regression line representing the expected quantitative behavior of the patterns' support).

These results suggest visually in the first read that there is no author that distinguishes himself considerably from the others in terms of quantitative employment of the syntactic patterns. Nonetheless, if we take a deeper looks on the highly ranked patterns (those extracted as relevant for describing the syntactic style of a given novel) and by analyzing more closely the textual instances for each one of them, we will notice that these results do not imply that the authors are not showing a distinguished syntactically marked language. Indeed, using the proposed method, the extracted patterns seem to have a strong relevance to characterize not only the style of the authors of our corpus but also to the novels' content and the literary genre in which it operates.

In what follows, some individual patterns among those extracted for each novel are discussed.

In Flaubert's *Madame Bovary*, several extracted patterns represent accurately the rhythmic rather than functional role of punctuation that is peculiar to the style of Flaubert. For example Pattern (1) captures instances of a comma preceding the conjunction, followed by a parenthetical clause:

Pattern (1): <(PUN) (KON) (PUN) (PRP)>, with support= 113, sample instances of the pattern in the text:

- , et , à
- , mais , avant
- ; et , à

This is actually a very peculiar property characterizing the individual style of Flaubert. This property was already manually identified and pointed out by Mangiapane (2012), but what is special about it in this case is that, using the proposed knowledge discovery process based on the present interestingness measures, we were able to automatically identify it without any prior knowledge about Flaubert's style, and this was done just by analyzing the statistical properties of the extracted patterns that contains it.

In *Le Ventre de Paris* by Zola, and in the same vein, the syntactic patterns extracted as relevant clearly represent the use of nested clauses to describe situations or attitudes in the novel such as in the Pattern (2), or to describe public places and objects in displays in long lists as in the Pattern (3):

Pattern (2): <(PUN) (PRP) (PRP) (NOM)>, support= 104, sample instances of the pattern in the text (bold text):

[23] Go back to Subsection 3.3.2 of Chapter 3 to know more about the α *ratio*

"Florent se heurtait à mille obstacles **, à des porteurs** qui se chargeaient **, à des marchandes** qui discutaient de leurs voix rudes ; il glissait sur le lit épais d' épluchures et de trognons qui couvrait la chaussée , il étouffait dans l' odeur puissante des feuilles écrasées ."

Pattern (3): <(NOM) (PUN) (PRP) (NOM) (ADJ)>, support= 68, sample instances of the pattern in the text:

- angles , à fenêtres étroites
- très-jolies , des légendes miraculeuses
- écrevisses , des nappes mouvantes

In *Eugénie Grandet* by Balzac, other different communicative functions are performed by the syntactic patterns ranked as relevant and their textual instances, for example:

Pattern (4): <(PUN) (VER) (NAM) (PRP)> , support= 49, is used as post-introducer of direct speech. This rather formulaic way of specifying (in a parenthetical form) the utterer of a reported speech is common to all, but seems to be strongly preferred by Balzac, while the other authors have shown a more varied style in introducing dialogues. Sample instances of the pattern in the novel:

- , dit Grandet en
- , reprit Charles en
- , dit Cruchot en

Pattern (5): <(NUM) (NUM) (NOM)>, support= 54, is a pattern used to refer to money, which is typical for the scenario of a novel in which money plays a very important role. Sample instances of the pattern in the novel:

- vingt mille francs
- deux mille louis
- sept mille livres

Pattern (6): <(ADV) (VER) (PRO) (ADV)>, support= 59, is used to express negative questions:

- n' avait -il pas
- ne disait -on pas
- ne serait -il pas

Pattern (7): <(PUN) (NOM) (PUN) (VER)>, support= 44, represents the punctuation extensively used to mimic spoken intonation and even to reproduce performance phenomena such as stutter:

- , messieurs , cria
- , madame , répondi
- , mademoiselle , disait

4.1.3. Correspondence Analysis Results and Discussion

In this subsection, we present some examples of relevant syntactic patterns extracted from the analyzed corpus using correspondence analysis-based measure along with a visualization analysis.

To start with, Figure 26 illustrates the plot resulting from the correspondence analysis projection of the four novels represented as vectors of patterns' supports. Novels (here labeled with the names of their authors as well) are diverging on the four axes, where patterns are unlabeled and printed in grey triangles with partial transparency.

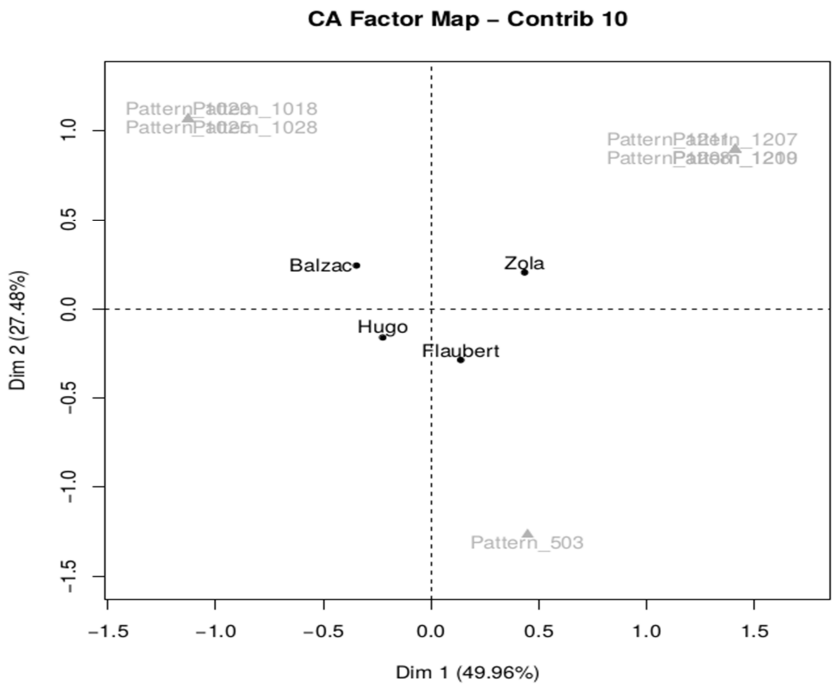

Figure 26. Top 10 most contributive patterns resulting from CA

As we can see, the analytic results confirm the intuition that we can have from the plot in Figure 26. Zola and Balzac are associated with the most contributive patterns, namely with patterns that are strongly over-used in their respective novels. Among the top 10 most contributive patterns in the plot only one is associated with Flaubert and none with Hugo. In fact, the first five patterns in order of contribution that stand closer to Hugo are ranked 80 to 339, while all other novels have associated patterns in the first 10 positions.

Is it possible to say that Balzac and Zola show a more syntactically marked language? In order to do that, we need to analyze more closely the instances for each pattern, and see if the differences in the pattern frequencies are due to stylistic reasons or to other more epiphenomenal facts.

In what follows some individual patterns among those extracted for each novel are discussed.

Zola

Pattern (8): `<(NOM) (ADJ) (PUN) (DET) (NOM)>` is very distinctive of Zola's *Le Ventre de Paris*. It occurs 188 times in this novel and close to none in the other ones. Typical instances of this pattern are sentences like:

- [Pattern (8_A)] "`Elle parut l' âme , la clarté vivante , l' idole saine et solide de la charcuterie ; et on ne la nomma plus que la belle Lisa .`"
- [Pattern (8_B)] "`Il était venu de Vernon sans manger , avec des rages et des désespoirs brusques qui le poussaient à mâcher les feuilles des haies qu' il longeait ; et il continuait à marcher , pris de crampes et de souleurs , le ventre plié , la vue troublée , les pieds comme tirés , sans qu' il en eût conscience , par cette image de Paris , au loin , très -loin , derrière l' horizon , qui l' appelait , qui l' attendait .`"

From the analysis of such instances it becomes clear that this pattern is used in descriptions and enumerations (8_A); it is also very used in parenthetical phrases (8_B) with the function of free adjuncts with adverbial function, namely modifying the verb (here "marcher", walk). Such phrases could be rewritten as normal prepositional phrases introduced by "avec" (with), but the author shows a strong preference for this structure.

The same can be said of Pattern (9): `<(DET) (NOM) (ADJ) (PUN) (DET)>` and Pattern (10): `<(PUN) (DET) (NOM) (ADJ) (PUN)>`, which is often a variation of Pattern (8).

Also Pattern (11): `<(VER) (PUN) (VER) (PRP)>` seems to be an expression of the same preference of Zola for implicit clauses to modify the verb and express manner.

Notice how all these patterns contain punctuation elements, often commas. The style of Zola is effective, with frequent use of parentheticals rather than explicit forms.

- [Pattern (11)] "`Il marchait , dormant à demi , dodelinant des oreilles , lorsque , à la hauteur de la rue de Longchamp , un sursaut de peur le planta net sur ses quatre pieds .`"

Instead, the second most important pattern for *Le ventre de Paris* – Pattern (12): `<(DET) (ADJ) (NAM)>` is associated with a very specific linguistic structure, namely with the modification of proper names, mostly of women. Here the feature identified seems more lexical than syntactical; probably Zola is trying to recreate the jargon of the Parisian populace, with people often being called by nicknames:

- `la petite Pauline`
- `la belle Normande`

Balzac

A first look at *Eugenie Grandet*'s patterns tells us that Balzac has a somewhat different style, with a preference for verbal structures and preposition, thus of explicit structures rather than implicit ones.

The first pattern is strongly associated with dialogical structures, which are very frequent in this work, Pattern (13): <(VER) (NAM) (PRP)>

- dit Grandet en
- reprit Charles en
- dit Eugénie en

The same can be said of Pattern (14): <(PUN) (VER) (PRO) (PRP)>, which is used mostly to (post-)introduce direct speech:

- "Bonjour , Grandet **, dit -il au** vigneron"
- "Mademoiselle **, dit -il à** Eugénie"

Pattern (15): <(NOM) (PRP) (VER) (DET) (NOM)> is associated with two structures, both verb phrases, with an explicative value (15_A) or to describe co-occurring events (15_B).

- [Pattern (15_A)] "Depuis le classement de ses différents clos , ses vignes , grâce à des soins constants , étaient devenues la tête du pays , mot technique en **usage pour indiquer les vignobles** qui produisent la première qualité de vin ."
- [Pattern (15_B)] "A cette observation , le notaire et le président dirent des mots plus ou moins malicieux ; mais l' abbé les regarda d' un air fin et résuma leurs pensées **en prenant une pincée de tabac** , et offrant sa tabatière à la ronde: Qui mieux que madame , dit -il , pourrait faire à monsieur les honneurs de Saumur ?"

Pattern (16): <(PRP) (NAM) (VER)> is used in phrases containing proper names, often place names in the function of modifiers.

- "L' Histoire **de France est** là tout entière ."
- "Les habitants **de Saumur étant** peu révolutionnaires ,..."

Pattern (17): <(VER) (DET) (NOM) (PRP) (VER)> shows a main transitive verb with its object and an implicit subordinate phrase. Like Pattern (15), it is used to better specify actions or events. Notice that basically this type of patterns constitutes the counterpart to those used by Zola, who prefers the verbless forms of predicate modification:

- "Charles **tendit la main en défaisant** son anneau"
- "Grandet **regarda sa fille sans trouver** un mot à dire ."

Thus Balzac's style is more verbose and explicit. The use of preposition to introduce phrases or clauses is important to highlight the relationship between head and modifier. Thus, it makes sentences less difficult to interpret. Balzac is considered the father of realism, but he aimed for a broader and more popular audience than Zola's, (for financial reasons as well as for artistic ones in our opinion). His style reflects possibly this necessity, as well as the time constraints of his immense production.

Flaubert

All of *Madame Bovary*'s patterns contain punctuation. The top five patterns all capture the same phenomenon, notably the fact that Flaubert's punctuation allows the comma to intervene before the conjunction as in:

- "Le soir, quand Charles rentrait, elle sortait de dessous ses draps ses longs bras maigres , les lui passait **autour du cou , et , l' ayant fait** asseoir au bord du lit , se mettait à lui parler de ses chagrins : il l' oubliait , il en aimait une autre !"

Patterns concerning punctuation style should always be taken with a pinch of salt, since punctuation in the edited version does not always reflect the choice of the author, but may be submitted to editorial choices. Nevertheless, as mentioned before, Mangiapane (2012) highlights the rather over rhythmical than functional role that punctuation has in Flaubert. Indeed from the rhythmical point of view, in the given example the commas mark the breathing pauses that is present as well before than after the conjunction "et". In fact, concerning the punctuation of Flaubert, here we notice the same properties extracted by the previous method.

Hugo

As was said before, Hugo's work is less marked than others. The patterns that do show some overrepresentation in *Notre Dame de Paris* are simple syntactic structures rather than complex ones.

Two of these patterns (Pattern (18): `<(NOM) (KON) (DET) (NOM) (PRP)>` and (Pattern (19): `<KON) (PRO) (NOM>`) are absent in Zola and Balzac, but are shared with Flaubert. Pattern (18) is the longest. It seems to be used mostly in the descriptions of places, which are very rich in the historical novel of Hugo, and helps the reader to enter into the world of medieval Paris:

- "Au centre de la haute façade gothique du Palais , le grand escalier , sans relâche remonté et descendu par un double courant qui , après s' être brisé sous le perron intermédiaire , s' épandait à larges vagues sur ses deux pentes latérales , le grand escalier , dis -je , ruisselait incessamment dans la **place comme une cascade dans** un lac .."

Pattern (19): `<(KON) (PRO) (NOM)>` is often used in subordinate clauses that show preference for demonstrative adjectives to underline situations.

- "Ajoutons que Coppenole était du peuple , et **que ce public** qui l' entourait était du peuple"
- "Et songer **que ce peuple** avait été sur le point de se rebeller contre monsieur le bailli , par impatience d' entendre son ouvrage !"

Pattern (20): `<(PUN) (KON) (VER)>` and Pattern (21): `<(NOM) (PUN) (KON)>` are shared with other authors, though slightly overrepresented in Flaubert. Here too, the punctuation variant found in Flaubert emerges, though not as strongly:

- "Quasimodo se plaça devant le prêtre , fit jouer les muscles de ses poings athlétiques **, et regarda** les assaillants avec le grincement de dents d' un tigre fâché ."

Pattern (22): `<(ADV) (ADJ) (KON)>` finally is used in comparisons and descriptions:

- "`Qui est aussi fraîche et aussi gaie que si elle était veuve .`"

By this analysis, the style of Hugo seems to emerge as full of lively descriptions, simple, personal, engaging, and popular just as we know it from the literary tradition.

4.1.4. Distribution Peculiarity Results and Discussion

In this subsection, some of the significant patterns extracted and ranked in terms of relevancy with the proposed method in Victor Hugo's novel *Notre-Dame de Paris* (NDdP) are described and discussed. We decided to focus on that specific novel because as it was shown in the analysis of the results from the two previous measures, it seems that Hugo's work is less marked than others. Since Hugo's style is not so distinctive in the comparative studies, we use the present intrinsic measure (that does not apply a comparative schema) to try to analyze it.

Figure 27 illustrates the distribution phenomenon (clustering) that characterizes three highly ranked patterns (relevant syntactic patterns in right column) as opposed to three non-relevant patterns (left column).

In order to appreciate the results, we shall compare them with the mere study of frequent patterns (frequency as interestingness measure). Frequency-based approaches may be used in comparative works, as in such cases it is possible to filter out patterns that are frequent in all texts and thus uninteresting. In an intrinsic approach this is not possible, thus the most frequent patterns are not very informative per se:

For example, the pattern: `<(VER) (DET) (NOM)>` (support = 3101) is one of the most frequent ones in NDdP. By comparing NDdP with other novels by contemporary authors, we might find out that Hugo under/over uses this syntactic structure, and possibly draw some conclusions. But in isolation this doesn't tell us much, as it is clear that the verb-object structure is very common in French.

The same is true for the pattern: `<(NOM) (PRP) (NOM)>` (support = 3101) that presents a simple noun phrase modified by a prepositional phrase which is a frequent structure in French.

On the contrary, by using the proposed method the extracted patterns seem to bear a strong relation to this particular text, its story line and the literary genre it instantiates, namely that of the historical novel.

Let us here take into account some examples:

Pattern (23): `<(NOM) (PRP) (NAM) (PUN)>`, support= 340, instances in the text:

- `tour de Notre-Dame ,`
- `hôtel de Bourbon ,`
- `murailles de Paris ,`
- `prince de Conty :`
- `dauphin de Vienne ;`

In Pattern (23), the proper name is often a location, especially at the beginning of the novel where descriptive parts are more frequent for the purpose of guiding the reader into the topography of medieval Paris. Later it serves the purpose of locating the plot. Other instances of this pattern are used to mention characters, especially historical ones, by their title and provenance. This also is very typical of a literary genre where historical elements are combined with fictional ones.

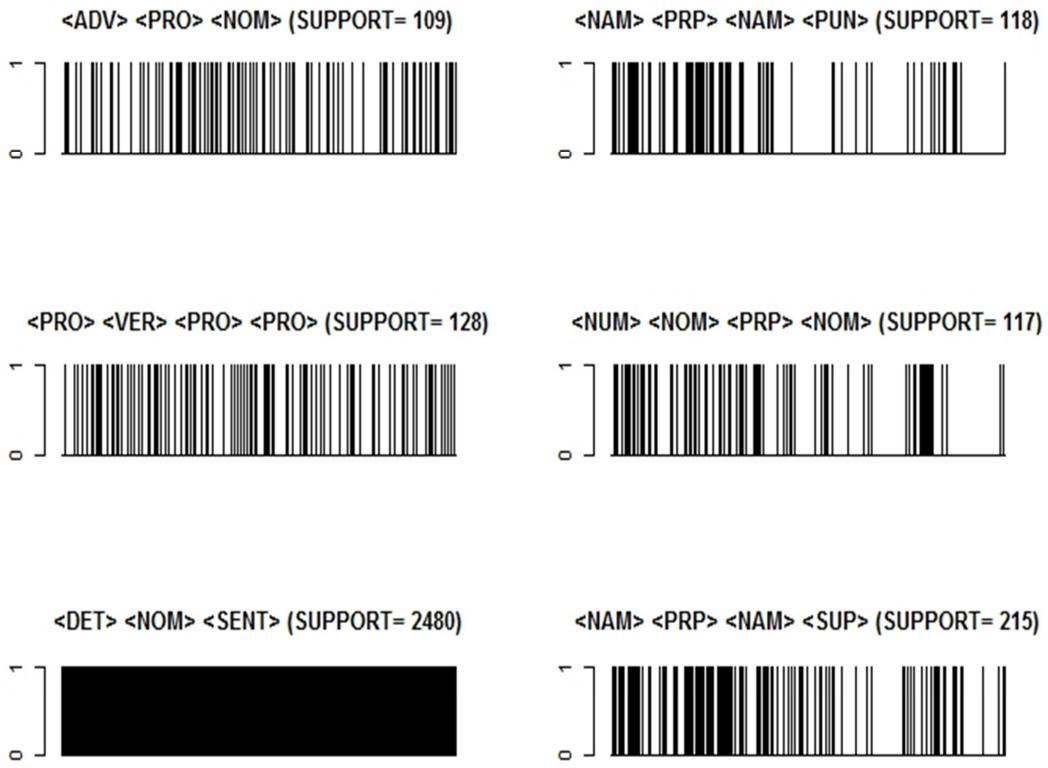

Figure 27. Positions' plot of three relevant syntactic patterns (right column) versus three non-relevant patterns (left column).

The skewedness of the distribution of this pattern is also due to the fact that some of the parts of the novel are more descriptive; they serve the purpose of introducing the historical setting and the characters, while other parts develop the action and thus do not introduce many new characters. Other extracted patterns have a similar function. Pattern (24) is often used to introduce characters by first stating their name and title. It is worth noticing that NDdP presents a plethora of minor characters (see positions plot of this pattern in Fig. 3):

Pattern (24): `<(NAM) (PRP) (NAM) (PUN)>` , support= 118, instances in the text:

- `Marguerite de Flandre ,`
- `Jehan de Troyes ,`

- `Frollo du Moulin ,`

The same is true for Pattern (25) which is also used to introduce character's full names (patronymics). Pattern (25): `<(PUN) (NOM) (NAM)>`, support= 120.

Pattern (26) instead is often instantiated in presentative structures such as (26_A and 26_B), with the topic (here the person) in focus position, used for changes of scenes, to introduce new characters.

In the final part of NDdP, this pattern instantiates other kinds of structures, such as (26_C), which are used to represent actions.

Pattern (26): `<(PRP) (VER) (NAM)>`, support= 113, instances in the text:

- [Pattern (26_A)] "`Il y avait pourtant une créature humaine que Quasimodo exceptait de sa malice et de sa haine pour les autres, et qu'il aimait autant, plus peut-être que sa cathédrale ; c'était Claude Frollo .`"
- [Pattern (26_B)] "**`C' était Quasimodo`**`, sanglé , cerclé , ficelé , garrotté et sous bonne garde .`"
- [Pattern (26_C)] "`…acculés à Notre-Dame qu' ils assaillaient encore et que défendait Quasimodo ,…`"

As we have seen, many of the significant patterns extracted with this technique contain the NAM (proper name) tag. This happens not only with sequential pattern mining, but also with other statistical pattern mining methods, as in general proper names are less frequent than other tags and their skewed distribution causes them to emerge in significance measurements. In a study that focuses purely on syntax, it may be worth merging this class with the one of common names.

Pattern (27) does not contain proper names, and seems very relevant for the text in question. Among the instances of this pattern we find many vivid and precise descriptions, as is evident especially in (27.B) where Hugo lists all the different divisions that used to be in charge of the defense of the former stronghold of "Châtelet", in Paris.

Pattern (27): `<(NUM) (NOM) (PRP) (NOM)>`, support= 117, instances in the text:

- [Pattern (27_A)] "`...le fracas de tous les gros doubles pétards de la Saint-Jean, la décharge de vingt arquebuses à croc, la détonation de cette fameuse serpentine de la Tour de Billy,`"
- [Pattern (27_B)] "`...les cent vingt sergents à cheval, les cent vingt sergents à verge, le chevalier du guet avec son guet, son sous-guet, son contre-guet et son arrière-guet ?`"

Finally, as a recapitulative remark, the few analyzed examples indicate that the presented technique is effective in extracting interesting syntactic patterns from a single text, and this seems particularly promising for the analysis of such texts that, for their characteristics or for historical reasons, cannot support a comparative study.

4.2. Quantitative Evaluation

In this part of the chapter, our aim is to have a numerical assessment about the discriminant power of the patterns that are extracted as characterizing the style. This is done on the basis that the stylistic choices of an author are in fact the elements that can allow us to distinguish his writings from others.

In fact the most important property that we want these patterns to have is to be able to describe and characterize the stylistic choices of a given author at a relatively high level, which makes them by the way capable of bringing meaningful information about this author's style. This is done without focusing on the discriminant power of those patterns.

As we have seen in the previous section of the chapter, by conducting a qualitative analysis of the extracted patterns, we had strong indications that the presented knowledge discovery process, especially for the method based on the correspondence analysis interestingness measure, is fairly effective in extracting interesting syntactic patterns for describing the style of a text.

We quantitatively evaluate in a larger corpus how much the extracted patterns are suited to differentiate the writing of an author from another one using a clustering algorithm. In that matter, the extracted patterns are used as vector features to describe the text. If the extracted patterns are effective in regrouping the writings of an author correctly, that can be considered as another clue of the stylistic relevance of these patterns. Actually, we have found this type of stylistic pattern to have a good performance in such task.

4.2.1. Experimental Settings

4.2.1.1. Analyzed Corpus

In this experiment, for the sake of consistency, we used novels written by the same four famous classic French authors as before, namely: Balzac, Flaubert, Hugo, and Zola. As explained before, this choice is motivated by the particular interest the Labex OBVIL has in studying the style of the classical French literature of the 19th century of which those four authors are well noticeable and prominent figures, except that this time we take four novels for each author. More details about the corpus used in this experiment are presented in Table 8.

4.2.1.2. Stylistic Features

Since we are evaluating the discriminant power of the characterizing patterns identified via the three proposed interestingness measures, these patterns are used as features to describe each text among others that include the most frequent patterns and some other stylistics features (used for a comparative purpose). This is done in such a way that each text is represented as a vector of supports of the syntactic patterns. Each text is represented as different frequencies vectors of the remaining stylistics features as well. In addition to that, we decided to vary the size of the feature set for each style marker to see how this could affect the clustering performance. For instance, concerning the patterns related to the quantitative-peculiarity based interestingness measure, we took respectively the most relevant 50^{th}, 100^{th}, 200^{th}, 300^{th}, 400^{th}, 500^{th} patterns as representative for each single text. We did the same for the remaining features. The full set of considered stylistic features is presented in Table 9.

4.2.1.3. Clustering Schema and Evaluation Measures

To conduct this evaluation experiment, we use an agglomerative hierarchical clustering method that seeks to build a hierarchy of clusters in a bottom-up approach (each observation starts in its own cluster, and then pairs of clusters are merged based on some cluster similarity criteria as one. This process is repeated until the top of the hierarchy is hit. The resulting hierarchy structure is called a dendrogram. Many cluster similarity criteria have been proposed and studied in the literature. In this experiment, we used the most famous and effective one called *ward algorithm* (Ward Jr 1963).

Table 8. The analyzed corpus for the quantitative evaluation

Author	Work	# of word	Label
Balzac, Honoré de	*Eugénie Grandet*	62849	Balzac_eugenie-grandet
	La maison du chat qui pelote	20849	Balzac_la-maison-du-chat-qui-pelote
	Le médecin de campagne	85217	Balzac_le-medecin-de-campagne
	Lys dans la vallée	101883	Balzac_Lys-dans-la-vallee
Flaubert, Gustave	*Bouvard et Pécuchet*	86711	Flaubert_bouvard-et-pecuchet
	Madame Bovary	111109	Flaubert_madame-bovary
	Trois contes	29730	Flaubert_trois-contes
	Un cœur simple	11055	Flaubert_un-coeur-simple
Hugo, Victor	*les misérables tome-I*	108270	Hugo_les-miserables-tome-I
	les misérables tome- II	93714	Hugo_les-miserables-tome-II
	L'homme qui rit	192777	Hugo_lHomme-qui-rit
	Notre dame de paris	168624	Hugo_notre-dame-de-paris
Zola, Émile	*Au bonheur des dames*	147710	Zola_au-bonheur-des-dames
	Germinal	164488	Zola_germinal
	L'assommoir	158857	Zola_lAssommoir
	Le ventre de Paris	110558	Zola_le-ventre-de-Paris

Table 9. Stylistic features considered for the quantitative evaluation

Stylistic feature	**Description**
QP_n	The top n^{th} most relevant pattern (Quantitative-peculiarity based measure)
CA_n	The top n^{th} most relevant pattern (Correspondence-analysis based measure)
DP_n	The top n^{th} most relevant pattern (Distribution-peculiarity based measure)
Lemma_n	The top n^{th} most frequent lemma
MF_n	The top n^{th} most frequent pattern
POS_28	Frequency of the 28 POS tags

Basically, each author in the analyzed corpus is considered to be representing a class, that is to say that each four novels written by the same given author are considered to be data instances belonging to the same natural class labeled by the name of the corresponding author. Concretely, each novel is represented as a vector of supports of the syntactic patterns and also as a frequencies' vector of each one of the other stylistics markers considered in this experiment. Once regrouped by the clustering algorithm, we conduct an evaluation of the resulting hierarchy structure representing the clustering configuration.

Specifically, we wish to answer the following questions: Can we put the writings of each one of the four authors in a separate cluster (four different clusters, each cluster containing the novels of one specific author)? How much are the stylistic patterns effective in doing so if we split the dendrogram into exactly four clusters?

To do so, we use two different cluster analysis evaluation indexes to evaluate the behavior and performance of the produced clusters. These two indexes (measures) evaluate the clustering configuration in two different ways: from the uniformity standpoint of the produced clusters using the Gini index, and also from the viewpoint of the accuracy of the decisions taken by the clustering algorithm using the Rand index.

Gini index

The Gini index (Farris 2010) is a measure of the degree of inequality in a data distribution. In our case, each class of the distribution produced by the clustering algorithm will be considered a distribution.

The Gini index takes a value ranging from 0 to 1, where 0 means perfect homogeneity (all data actually belong to the same natural class) and 1 means a total heterogeneity. With contextualization, it is calculated as follows:

Let C_k be an automatically produced cluster, its corresponding Gini index $Gini_{C_k}$ is computed as follows:

$$Gini_{C_k} = 1 - \sum_{i=1}^{n} p_i^2$$

Where:

- n is the number of authors (natural classes) appearing in C_k
- p_i is the number of novels (data instances) written by author i in C_k

So, the overall Gini index of the resulting clustering configuration $Gini_{overall}$ is:

$$Gini_{overall} = \sum_{k=1}^{K} \frac{N_k=4}{N=16} Gini_{C_k}$$

Where:

- N total number of novels (data instances) which is equal to 16
- N_k number of novels for each author which is equal to 4
- K is equal to 4 since we are splitting the dendrogram into 4 clusters

Rand index

The Rand index (Rand 1971) is a statistical index that measures the similarity between two data groups (repartitions). In our case, these distributions will be the distribution of the natural classif-

cation in which each novel belongs to its author's class on the one hand, and the distribution produced by the clustering algorithm on the other hand. From a mathematical point of view, the Rand index is quite similar to the accuracy measure well known in the evaluation of supervised machine learning algorithm. Rand Index is computed as follows:

$$Rand = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP is the number of novels' pair written by the same author and put in the same cluster by the clustering algorithm
- TN is the number of novels' pair written by different authors and put in different clusters by the clustering algorithm
- FP is the number of novels' pair written by the same author and put in different clusters by the clustering algorithm
- FN is the number of novels' pair written by different authors and put in the same cluster by the clustering algorithm

4.2.2. Results and Analysis

Table 10 and 11 illustrate the results of the evaluation of the clustering algorithm for the different style markers using both Gini and Rand index. Table 10 illustrate the best top 5 performing features sorted according to the Rand index value, while Table 11 contains the full set.

Table 10. Top 5 performing features sorted according to Rand index

Style markers	**Gini index**	**Rand index (%)**
CA_300	0,25	85,83
CA_200	0,29	83,33
CA_400	0,31	82,50
QP_150	0,31	82,50
CA_100	0,35	81,66

Table 11. Results of the evaluation of the clustering algorithm for the different stylistic features

Style markers	**Gini index**	**Rand index (%)**
QP_50	0,44	67,50
QP_100	0,47	58,33
QP_150	0,31	82,50
QP_200	0,46	68,33
QP_300	0,44	67,50
QP_400	0,45	70,83
QP_500	0,44	69,16
CA_50	0,46	68,33
CA_100	0,35	81,66
CA_150	0,44	67,50
CA_200	0,29	83,33
CA_300	**0,25**	**85,83**
CA_400	0,31	82,50
CA_500	0,34	80,83
DP_50	0,44	67,50
DP_100	0,48	56,66
DP_150	0,47	58,33
DP_200	0,54	60,83
DP_300	0,44	67,50
DP_400	0,55	69,16
DP_500	0,54	69,16
Lemma_20	0,47	72,50
Lemma_30	0,44	67,50
Lemma_50	0,34	80,83
Lemma_100	0,36	72,50
Lemma_200	0,39	75,83
Lemma_300	0,36	72,50
Lemma_500	0,46	70,83
Lemma_1000	0,45	70,83
Lemma_5000	0,52	70,83
MF_50	0,57	69,16
MF_100	0,54	69,16
MF_150	0,54	69,16
MF_200	0,54	69,16
MF_300	0,44	67,50
MF_400	0,45	70,83
MF_500	0,37	77,50
POS_28	0,54	69,16

The first thing that we notice from these results is the nearly perfect correlation between the two clustering evaluation indexes results. The best performing features in terms of Rand index measures are also the best performing in terms of Gini index and vice versa with few exceptions.

Clearly, patterns related to correspondence analysis interestingness measures are outperforming all other features including the patterns related to the other two interestingness measures. This is actually true in both the accuracy of clustering decision and the homogeneity of the produced clusters (the most 300^{th} relevant patterns). This is somehow expected since the statistical property of the correspondence analysis technique, which is a cross-comparative technique, gives its corresponding patterns an advantage in terms of discriminant power. In terms of performance, the correspondence analysis-based patterns are followed by other features, respectively by the most 150^{th} relevant patterns related to the quantitative-peculiarity interestingness measure, the most 50^{th} frequent Lemma and the most 500^{th} frequent syntactic patterns (without any interestingness relevancy).

Increasing the support threshold of the style marker, to take into account individually in the clustering experiment, does not necessarily increase or decrease the clustering discriminant power and consequently the clustering performance (CA_200 is performing better than CA_400 but worse than CA_300). However such behavior can give us an idea about the empirical optimum threshold to be taken into account if one should limit its analysis to a restricted set of patterns.

The pattern ranked as relevant by the distribution-peculiarity based measure are particularly performing poorly on this clustering task. We think that this is due to the fact that the technical formulation of this measure is biased toward the extraction of patterns that are locally relevant in each text and thus not much suitable for characterizing it in its integrity.

The produced hierarchical clustering structure for the most 300^{th} relevant syntactic patterns of the correspondence analysis interestingness measures (best performing features) is illustrated in the Figure 28's dendrogram. As we can see, based on these patterns the algorithm was successfully able to isolate all the novels written by both Zola and Flaubert in separated clusters. Balzac's novel *la maison du chat qui pelote* was however mistakenly regrouped within the Flaubert's cluster. In another side, the clustering algorithm was less effective in separating the writings of Balzac and Hugo, especially for the two parts of the novel *les miserables* by Hugo which were both regrouped with Balzac's works. Remarkably, this goes in line with what we have noticed in the correspondence analysis projection presented in Figure 26 in which Zola is associated partly with the most contributive patterns while Hugo is shown to be less syntactically remarkable.

We think that this is another strong indication that the contribution power can play the role of an interestingness measure for identifying the most stylistically relevant syntactic patterns.

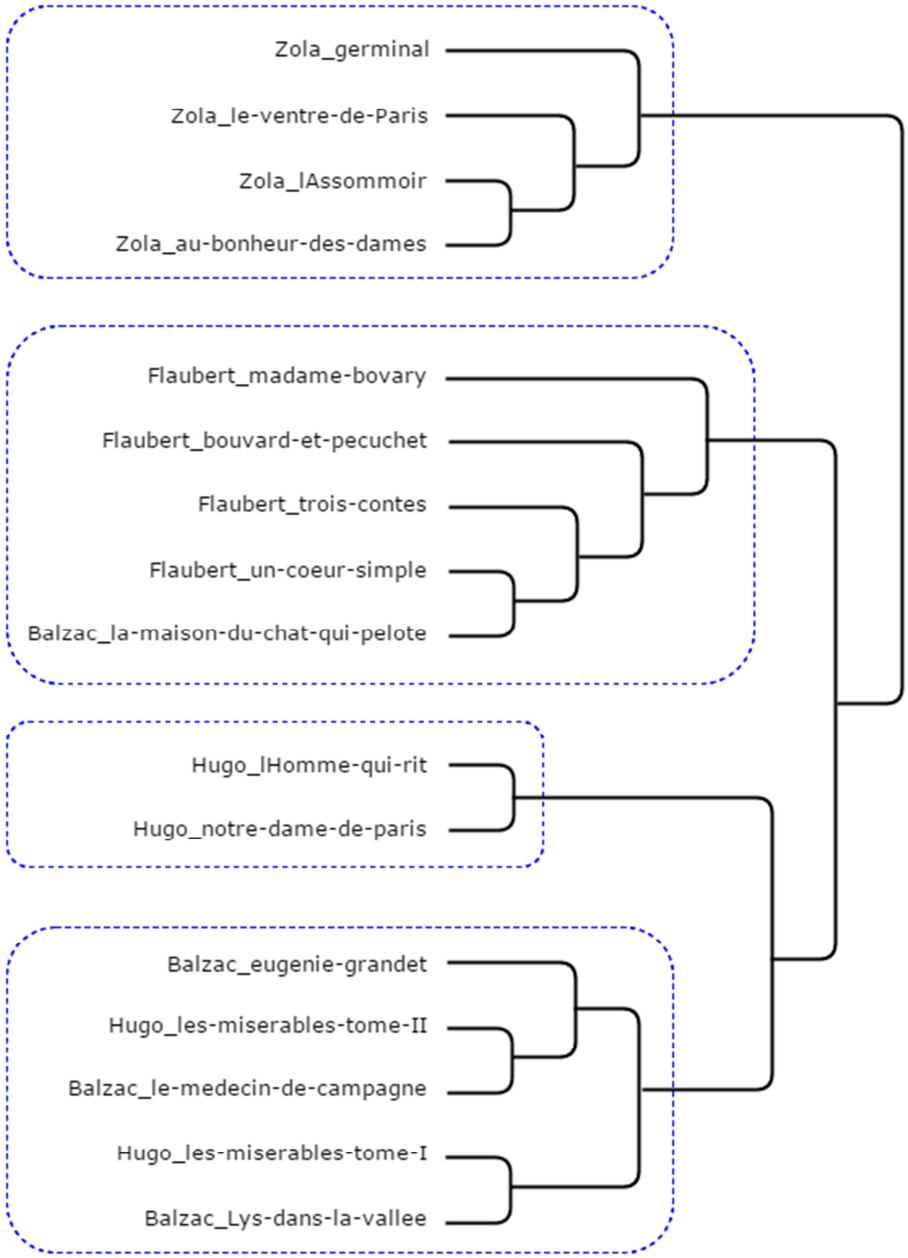

Figure 28. Hierarchical clustering structure for correspondence analysis top 300th patterns

4.2.3. General Discussion

The few analyzed examples indicate that the presented techniques are fairly effective in extracting interesting syntactic patterns, and this seems particularly promising as a computer-assisted literary analysis tool to support linguists and literary researchers in their critic analysis.

Of course, these results are far from being perfect. In fact the critical samples of syntactic patterns presented in the qualitative analysis part of this chapter do not represent the overall quality of the

extracted patterns. There exist patterns that do not show necessarily neither strong nor relevant relations to any kind of identified stylistic preferences.

From an internal comparative point of view, it is clear that the correspondence analysis-based measure was the best performing method both in terms of qualitative and quantitative results. This makes it the more suitable method for dealing with such tasks for many reasons. (Indeed, we chose to work on this method in the case study presented in the next chapter).

Actually, to begin with, the visualization aspect of this method is very intuitive. Indeed, it is quite possible for a literary researcher unfamiliar with everything related to computation and automatic methods to imagine a mapping between the notions of proximity in the two dimensional space employed in the correspondence analysis and the notion of characterization. This is definitely more expressive than some numerical values evaluating this property. This can also considerably help them to more easily understand the results. In fact, we found out that the researchers from Labex OBVIl, with whom we have worked, have particularly appreciated the visualization aspect offered by this method compared to the other ones.

The second point which can be counted on the plus side for the correspondence analysis-based measure against the other ones is the general analysis view that it offers. Actually, this method permits to include all the analyzed texts in one single process and to put them in one unified illustration. This can help users to produce a more clear comparative analysis with respect to what we can have when the patterns are assessed for each text separately (as done in the quantitative peculiarity and the quantitative distribution-based methods).

The third point that makes the correspondence analysis–based measure better than the other measures is the property that has its resulting patterns in terms of their ability to discriminate some author's works from others. Actually, in addition to be fairly characterizing the style of a given author, these patterns have a fairly discriminant power, despite the fact that they belong to relatively high linguistic level of description. These patterns were able to outperform in terms of clustering performance other style markers known to be highly effective in such a task.

Finally, by doing a comparative quantitative analysis of the patterns related to the three proposed interestingness measures, we found out that the correspondence analysis-based measure is the most consistent one in terms of produced results[24]. That is to say the measure was able to identify more or less the same patterns for different works written by the same author. Considering that the style of an author is composed of a consistent set of stylistics traits and choices, this can be seen as another proof of the viability of this measure.

[24] The consistency was evaluated using the overlapping between the set of relevant patterns resulting from different novels written by the same author. The Jaccard distance was used as an overlapping measure. For instance, the average Jaccard overlapping value for the correspondence analysis-based results was equal to 0.79 which is quite good knowing that Jaccard distance value range from 0 (total msimatch) to 1 (total overlapping)

Chapter 5. Studying the Stylistic Characterization of Molière's Characters

The work presented in this chapter, including the research questions and the result analysis and discussion, was performed with the close and valuable collaboration of Dr. Francesca Frontini, researcher at Istituto di Linguistica Computazionale in Pisa (Former Postdoc at Labex OBVIL in Paris) and Dr. Elodie Benard, Posdoc at Labex OBVIL in Paris, and with the appreciated advices and suggestions of Prof. Georges Forestier.

Successful writers of literary fiction and theatre plays are generally renowned for their ability to create memorable characters that take on a life of their own and become almost as real as living people for their readers/audiences. The study of theatrical characterization, namely the investigation into how these effects are achieved, is not a new topic in computational stylistics or in corpus studies. Mahlberg (2013) attempts to identify typical lexical patterns for memorable Dickens' characters by extracting those lexical bundles that stand out (namely those that are over-represented) in comparison with those found in a more general corpus. As explained previously in Chapter 2 (Literature Review on Computational Stylistics), such methods represent a more hermeneutical approach in that they offer literary criticism a powerful tool for interpretation, while not trying to replace the insight of the human. Other works, that are more in the line of classification approach, apply authorship attribution methods to the different characters of a play to identify whether the author has managed to provide each of them with a stylistically distinct voice. For instance (Vogel & Lynch 2008) compare the dialogue of individual Shakespearean protagonists against the text of the whole of a play or even against all plays from the same

author. In most cases lexical or even sub-lexical elements (character n-grams) are used as features in the analysis.

In this chapter, we apply the proposed knowledge discovery process for studying the characterization in classic French plays from a syntactic point of view. The work we present here is intended to support syntactic textual analysis in two ways, namely by:

- Verifying the degree of characterization of each character with respect to others, and
- Automatically inducing a list of linguistic features that are significant and representative for that character

The methodology relies on the correspondence analysis-based interestingness measure for the comparison of pattern characterization for each character and for the visual representation of such differences. As we have seen in the previous chapter, this interestingness measure was shown to be the most effective in extracting relevant syntactic patterns from texts.

In this case study's chapter, we report on four different experiments conducted on the work of the so famous French playwright Molière[25], cross-comparing characters from different plays. Each experiment is more or less complementary to another. In the first experiment, we focus on the stylistic analysis of four memorable protagonist characters of prose plays written by Molière. Our aim in this first experiment is to study the stylistic singularity that Molière gives to his protagonists in its syntactic form. In the second experiment, we took another perspective; we conduct a comparative analysis of different characters that share the same name, namely *Sganarelle*, in Molière's plays. Our aim is to find out what is really common between these characters beside the name, and what stylistically differs between them. In the third experiment, we make a crossover between the first and the second experiment in the sense that we include in one single analysis both the four protagonists of the first experiment and Sganarelle's characters. Finally, in the fourth and last experiment (in which we address a quite different question), we focus on the figure of the "*raisonneurs*", characters who take part in discussions with comical protagonists providing a counterpart to their follies, in an attempt to identify the differences marked by the stylistic choices of Molière. For each experiment, we report on some of the most illustrative extracted morpho-syntactic patterns in terms of their capability to stylistically qualify the text under investigation.

5.1. First Experiment: Molière's Memorable Protagonists

In this first experiment[26] carried on Molière plays, we focus on a transversal study on different characters, namely four main characters of prose plays. Patterns were extracted from the texts of

[25] We use the edition of the Project Molière, supervised by Prof. Georges Forestier at the Labex OBVIL (http://obvil.paris-sorbonne.fr/projets/projet-moliere?equipe)

[26] In this first experiment (as well as in the fourth one) and as a matter of illustrating the different extraction options that can be taken into account in the analysis process, we decided to work on a different extraction option settings. Actually, we include in this analysis the possibility to have gaps (jokers) in the patterns. We work with an extended POS tag set that includes the morphological information as well

these four memorable Molière protagonists which have been extracted by separating them from the rest of their respective plays. They are:

- Scapin (*Les Fourberies de Scapin*)
- Sganarelle (*Le médecin malgré lui*)
- Harpagon (*Avare*)
- Dom Juan (*Dom Juan*)

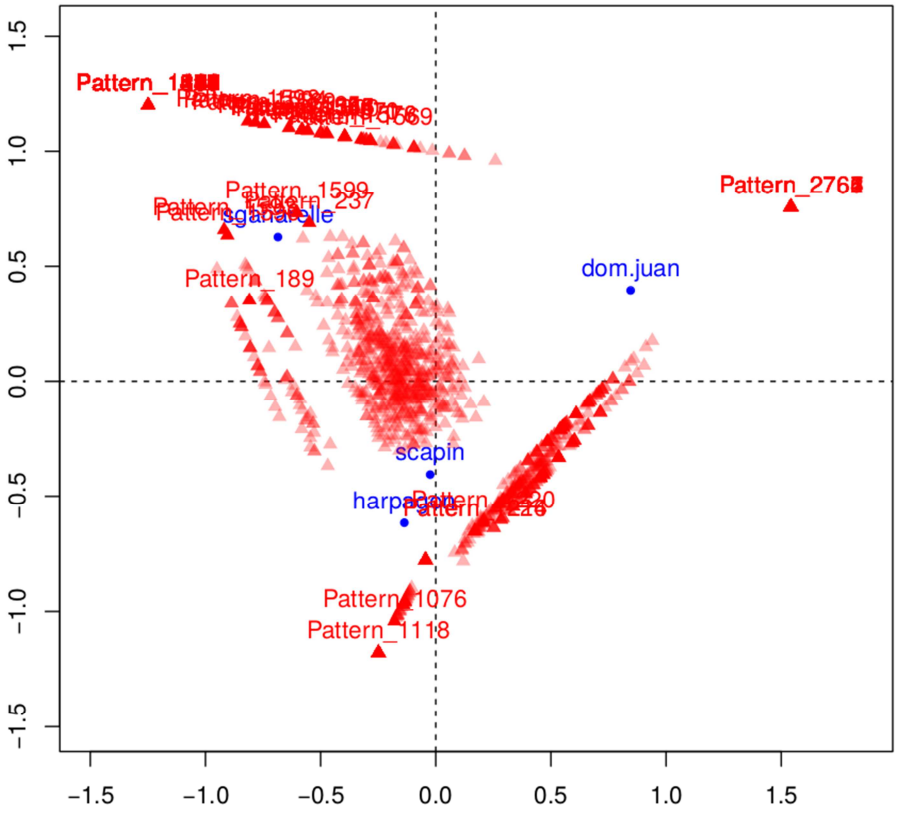

Figure 29. Correspondence analysis of four memorable Molière protagonists

The plot in Figure 29 shows the relative distances between the four characters according to correspondence analysis; all patterns are in shadow, except for the first 200 by contribution. The most isolated character seems to be Sganarelle, the protagonist of a piece in which a simple man is forced by circumstances to pretend to be a great doctor; that is, his language is quite different, from a syntactic point of view, from the other protagonists. Among his most significant patterns we find syntactic structures that are typically used to express diagnosis (Pattern 28).

Pattern (28): `<(PRO:PER) (VER:pres) (KON) (*) (NOM)>`[27] , instances of the pattern:

[27] '*' stand for a joker, which is a gap that can be filled with any (word lemma POS) itemset

- ".. **il arrive que ces vapeurs** ... Ossabandus , nequeys , nequer , potarinum , quipsa milus"
- "**je tiens que cet empêchement** de l' action de sa langue est causé par de certaines humeurs …"
- "**il se trouve que le poumon** , que nous appelons en latin armyan, …"
- "**on voit que l' inégalité** de leurs opinions dépend du mouvement oblique du cercle de la lune …"

In other cases such as in Pattern (29), the pattern groups assertions having a performative function and which are used initially to try and clear up misunderstandings (vainly it turns out), then to assure people of his assertions, and finally, once discovered, to confess.

Pattern (29): <(PRO:PER) (PRO:PER) (*) (KON)>

- je te dis que
- Je vous promets que
- Je vous jure que
- je vous dis que
- Je vous assure que
- je vous apprends que
- Je vous apprendrai que
- je vous avoue que

Dom Juan, a nobleman and a complex character, is instead isolated by under-representation, in that he has less distinctive patterns, which may mean that his language is less repetitive and, possibly more elaborate. This is also evident from one of the few patterns that are strongly associated with him (Pattern 30), which captures the over-use of subordinate clauses.

Pattern (30): <(KON) (PRO:PER) (*) (VER:pres) (*) (PRP)>

- "sachez **que je n' ai point d'**autre dessein que de vous épouser …"
- "elle va vous dire **que je lui ai promis de** l'épouser"
- "Vous soutenez également toutes deux **que je vous ai promis de** vous prendre pour femmes"
- "… et **que je sais me servir de** mon épée quand il le faut"

One should also take into consideration that the play *Dom Juan* was written by Molière in "prose rythmée**"** (rhythmic prose) which is not the case with the other plays in the current sample. This may also explain the isolation of this character given the higher degree of syntactic variability that metric constraints impose.

Finally the two comical characters Scapin and Harpagon are both characterized by patterns of lower syntactic complexity. This is especially the case with Harpagon (Patterns 31_A and 32) whose patterns convey the image of a self-centered person, who wants to have things his way, and who is subject to violent disappointments (especially when money is concerned).

Pattern (31_A): `<(PRO:PER) (PRO:PER) (VER:pres) (VER:pper)>`

- `on m' a privé`
- `on m' a dérobé`
- `on m' a volée`
- `on m' a pris`

Pattern (32): `<(KON) (PRO:PER) (*) (KON)>`

- `que je veux que`
- `et il faut que`
- `et vous verrez qu'`

These syntactic patterns have a slightly different function in Scapin, the clever servant who interacts with several characters in order to try to carry out his plan. In Pattern (31_B) we see the same pattern as in Pattern (31_A), but used mostly to report events.

Pattern (31_B): `<(PRO:PER) (PRO:PER) (VER:pres) (VER:pper)>`

- "**`Je l' ai trouvé`** `tantôt tout triste`"
- "**`nous nous sommes allés`** `promener sur le port.`"

It is worth noticing how such structures in the past tense are under-represented in the character of Sganarelle, whose discourse is prevalently in the present tense; while Dom Juan, Sganarelle and Scapin are all actively lying in their respective plots, the use of past tense in Scapin may be more reflective of conscious scheming.

5.2. Second Experiment: Molière's Sganarelles

As mentioned in the introduction of this chapter, in this second experiment we took another perspective by conducting a comparative analysis of different characters that share the same name, Sganarelle, in Molière's plays. An experiment was first conducted on the characters, in Molière's plays, which are called "Sganarelle". They appeared in seven comedies and can be distinguished in various ways.

Firstly, the social status:

- 4 are bourgeois (SGAcocu, SGAecole, SGAamour, SGAforce respectivally for *Le Cocu imaginaire*, *L'École des maris*, *L'Amour médecin* and *Le Mariage forcé*)
- 2 are servants (SGAjuan and SGAvol respectivally for *Dom Juan* and *Le Médecin volant*)

- 1 is a common man, a woodcutter (SGAmalgrelui for *Le Médecin malgré lui*)

Secondly, their function in the action:

- 2 pretend to be doctors (*Le Médecin malgré lui* and *Le médecin volant*): usually, disguise is used as a form of trickery, but as for Sganarelle in *Le Médecin malgré lui*, he fools "in spite of himself" and is fooled by his wife
- 4 are a jealous husband, a father who wants to marry her daughter against her will, and bachelors eager to get married; all are fooled
- 1 is Dom Juan's servant: Sganarelle dressed as a doctor, but it's a subsidiary incident for hiding his identity and it's above all a means to exhibit that Dom Juan is impious/heretic in medicine too

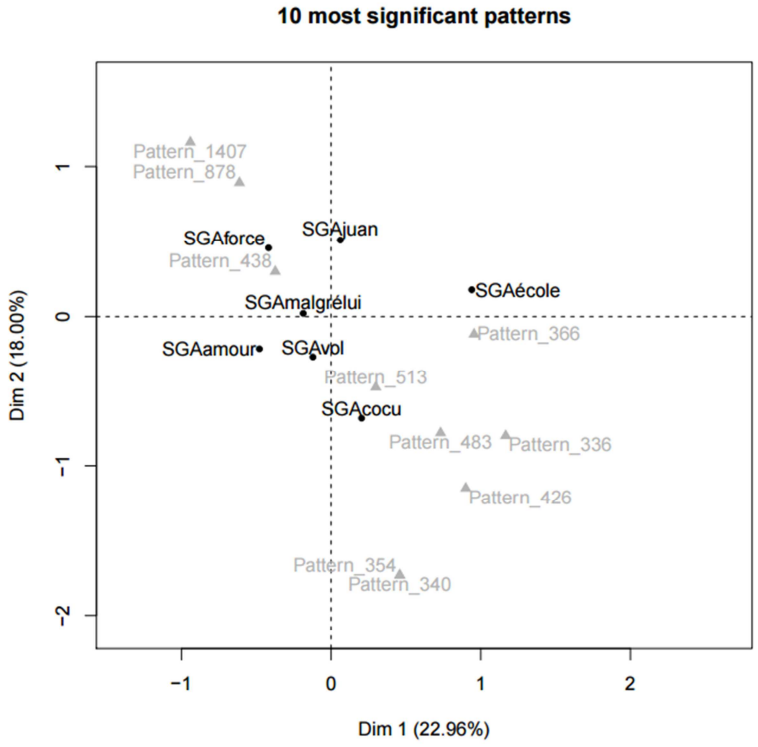

Figure 30. Correspondence analysis of the Sganarelles

The correspondence analysis plot in Figure 30 shows that Sganarelle from *L'École des maris*, who expresses himself in verse is relatively isolated on the right of the *x*-axis. However, the difference between prose form and verse form may not be meaningful because Sganarelle from *Le Cocu imaginaire*, which is also a play in verse, isn't as far on the right as Sganarelle from *L'École des maris*.

The plot doesn't seem to reflect the social difference between the bourgeois on the one hand, and the servants on the other. The characters appear to be rather located on the plot according to the kind of comic effects used by Molière.

5.2.1.1. Patterns related to comic effects

Patterns that seem to be highly distinctive are related to comic effects due to repetition. For example, two patterns extracted from Sganarelle from *L'Amour médecin* correspond to the same instances:

Pattern (33): `<(PRO) (PRO) (VER) (ADV) (SENT)>`, instances (6 in total):

- `Non, ne m'en parlez point.`
- `Ne m'en parlez point.`
- `Ne m'en parlez point.`

These 3 sentences are drawn from the scene 3 based on a dialogue of deaf[28] between Sganarelle and his daughter's servant, Lisette. Sganarelle doesn't want to hear Lisette, repeating that the only remedy for his daughter is a husband (Sganarelle don't want his daughter to get married for keeping his wealth).

Pattern (34): `<(PRO) (VER) (PRO) (SENT)>`, instances (8 in total):

- `Hé bien, qu'est-ce ?`
- `Que sera-ce ?`
- `Qu'est-ce ?`
- `Qu'y a-t-il ?`
- `Qu'est-ce ?`

These instances are taken from the scene 6 which is again a dialogue of deaf between Sganarelle and Lisette, but this time, it's Lisette who refuses to hear. She appears yelling "`Ah, Malheur! Ah disgrâce!`" and keeps shouting and crying without explaining to his master, who is getting more and more anxious, what has happened.

Pattern (35) associated with Sganarelle from *Le Mariage forcé* is also linked to a comic device:

Pattern (35): `<(PRO) (PUN) (NAM)>`

8 of the 9 instances refer to sentences left incomplete, containing the first person pronoun "je" and dots (for example: "**`Je … Eh!`**" where "`Eh!`" [29] is the first word of Sganarelle's next line). The comic effect lies on the fact that the Aristotelian philosopher Pancrace, only preoccupied by Aristotle's precepts, doesn't stop interrupting Sganarelle and preventing him from speaking.

It is interesting to note that the "medical comic", which characterizes *Le Médecin malgré lui* and *Le médecin volant*, both based on the comic device of the false doctor, isn't supported by repetitions but by the use of medical jargon and dog Latin, the prescription of absurd or dangerous remedies, the disagreement on the nature of the illness. Thus, the distinctive patterns associated with the Sganarelles of these plays aren't related to comic effects. Given that the

[28] Lisette: " On dit bien vrai ; qu'il n'y a point de pires sourds, que ceux qui ne veulent point entendre " (sc. 4) and Sganarelle: " Il est bon quelquefois de ne point trop faire semblant d'entendre les choses qu'on entend que trop bien " (sc. 5)

[29] As someone may notice, this is clearly a tokenization and POS tagging error, since that the textual instance does not syntactically correspond to Pattern (35). The correct corresponding pattern should be `<(PRO) (PUN) (INT) (SENT)>.` However, Pattern (35) still expresses its communication function correctly!

patterns related to a comic repetition are highly distinctive, it may explain that these two Sganarelles are never too far from the others characters. Indeed, we can note the relatively central position of these characters on the plot with respect to the others.

5.2.1.2. Patterns revealing characters traits

Some patterns can reveal character, as shown by the following 4 patterns.

Sganarelle from *Le Mariage forcé* is strongly associated with a pattern containing a verb referring to the act of seeking advice "conseiller, communiquer, écouter, répondre" preceded by an object pronoun which represents the speaker, taking advice or the listener. Indeed, like Panurge in *Le Tiers Livre*, Sganarelle desesperately searches for authority that will tell him whether he must marry.

Pattern (36): <(PRO) (VER) (SENT)>, instances:

- "Tout de bon, vous me **le conseillez ?**"
- "J'ai quelque chose à **vous communiquer.**"
- "La phrase : Voilà qui est fait : je vous prie de **m'écouter.**"
- "Laissez tout cela, et prenez la peine de **m'écouter.**"
- "Ce n'est pas là **me répondre.**"

A bit more complex pattern (overlapping with pattern (28) from the first experiment) distinguishes the Sganarelle from *Le Médecin malgré lui*. It contains a verb expressing certainty and confidence ("vouloir, jurer, assurer, apprendre" and modals) and followed by a that-clause, eloquent of the speaker's authority over his wife (reminding her who commands and threatening her), his prospective buyers, his patients (asserting diagnosis).

Pattern (37): <(PRO) (VER) (KON) (PRO)>, instances:

- "Non, je **te dis que je** n'en veux rien faire, et que c'est à moi de parler et d'être le maître."
- "**Il suffit que nous** savons ce que nous savons, et que tu fus bien heureuse de me trouver."
- "Ma femme, **vous savez que je** n'ai pas l'âme endurante, et que j'ai le bras assez bon."
- "Je **vous promets que je** ne saurais les donner à moins."
- "Si vous savez les choses, **vous savez que je** les vends cela."
- "Vous en pourrez trouver autre part à moins : il y a fagots et fagots ; mais pour ceux que je fais... Je **vous jure que vous** ne les auriez pas, s'il s'en fallait un double."
- "Je **vous assure que c'**est du meilleur de mon âme que je vous parle."
- "Je **vous assure que je** suis ravi que vous soyez unis ensemble."

- "Mais comme je m'intéresse à toute votre famille, **il faut que j'**essaye un peu le lait de votre nourrice, et que je visite son sein."

Sganarelle is well and truly that domestic tyrant which values run counter to those of the *mondains*, who advocate, on the contrary, the male submission[30]. The extraction highlights the parallel between sentences such as "Je tiens que cet empêchement…" and "Tous nos meilleurs auteurs vous diront que c'est l'empêchement" which signals that Sganarelle is given (or gives himself) equivalent status to the ancients.

As for Sganarelle from *Le Cocu imaginaire*, we find among his most significant patterns a syntactic structure that betrays the theme of credulity, which Sganarelle is an emblem of. Pattern (38) draws attention to the use of circumstantial complements referring to a body part for proving reality.

Pattern 38: <(VER) (PRP) (DET) (NOM) (VER)>, instances (6 in total):

- "À d'autres je vous prie, la chose est avérée, et je **tiens dans mes mains. Un** bon certificat du mal dont je me plains."
- "Ce Damoiseau, parlant par révérence me fait cocu Madame, avec toute licence ; et j'ai **su par mes yeux avérer** aujourd'hui le commerce secret de ma femme et de lui."
- "Sans doute, et je l'**avais de ses mains arraché**, et n'eusse pas sans lui découvert son péché."

Finally, in the discourse of Sganarelle from *Dom Juan*, one of the distinctive patterns refers to negative rhetorical questions, which are used in contexts where the character must argue and convince. Sganarelle keeps trying, throughout the comedy, to convince Dom Juan that God exists and will take his revenge on him.

Pattern (39): <(ADV) (VER) (PRO) (ADV)>, instances (6 in total):

- "**Ne voyez-vous pas** bien, dès qu'on en prend, de quelle manière obligeante on en use avec tout le monde, et comme on est ravi d'en donner à droit et à gauche, partout où l'on se trouve ?"
- "Osez-vous bien ainsi vous jouer au Ciel, et **ne tremblez-vous point** de vous moquer comme vous faites des choses les plus saintes ?"
- "**Ne croyez-vous point** l'autre vie ?"
- "Vous voilà vous, par exemple, vous êtes là : est-ce que vous vous êtes fait tout seul, et **n'a-t-il pas** fallu que votre père ait engrossé votre mère pour vous faire ?"
- "Cela **n'est-il pas** merveilleux que me voilà ici, et que j'aie quelque chose dans la tête qui pense cent choses différentes en un moment, et fait de mon corps tout ce qu'elle veut ?"

[30] Georges Forestier and Claude Bourqui, Preface, in Molière, *Œuvres complètes*, Paris: Gallimard, t. I, p. XXX and XLII-XLIII.

- "**Ne sais-je pas** bien que je vous dois ?"

5.3. Third Experiment: Molière's Protagonists vs. Molière's Sganarelles

In the third experiment, we include in one single analysis both the four protagonists of the first experiment and Sganarelle's characters. By the mean of this experiment, we intend to put insight on some research questions and ideas:

- Are Sganarelles clustering with each other?
- Are they similar or not to other characters?
- Forestier hypothesis: Sganarelle of *Médecin Volant* is less prototypical with respect to the other ones

In fact concerning the third question, Georges Forestier & Bourqui (2010) said that: "*Il est par ailleurs possible, même s'il s'agit de l'une des premières pièces écrites par Molière, que le texte qui nous en est parvenu soit celui d'une version remaniée plus tardivement. C'est ce qui expliquerait notamment le choix étonnant de donner le nom de Sganarelle au héros du Médecin volant, alors que son caractère ressemble beaucoup à celui du zanni (le valet rusé) de la commedia dell'arte, qui dans les premières pièces de Molière apparait généralement sous le nom de Mascarille. Dans cette hypothèse, il s'agirait d'une version de la pièce retravaillée pour les quatre représentations de 1664, à une époque où les personnages qu'incarnait Molière au théâtre (hormis dans les « grandes comédies ») portaient généralement le nom de Sganarelle*" [31].

Notice in Figure 31, representing the correspondence analysis projection of both four memorable Molière's protagonists and the Sganarelles, how the plays in poetry are much more distant in terms of syntactic features than in terms of lexicon. Sganarelles of *Ecole des Maris* and of *Cocu Imaginaire* cluster on the left right of the screen in the correspondence analysis plot.

In this first figure, you can see the clear horizontal distinction between the prose and the poetry characters. This is a confirmation of what we saw in other experiments; syntax is very sensitive to genre and register variation which makes it very difficult to study.

We compare for instance with this other plot, generated with another tool for stylometry, namely stylo for R[32]. Here a similar analysis is carried out using word frequencies as features instead of syntactic patterns.

[31] "It is also possible, although it is one of the first plays written by Molière, that the text that reached us is that of a lately reworked version. This particularly explains the surprising choice of giving the name of Sganarelle to the hero of the flying doctor, while his character is much like that of zanni (the wily servant) of the commedia dell'arte, which in the early plays of Molière generally appears as Mascarille. In this hypothesis, it would be a reworked version of the play for the four performances of 1664, at a time when the characters embodied by Molière himself in theatre (except the "great comedies") generally were named Sganarelle" [translation provided by the thesis' author]

[32] https://sites.google.com/site/computationalstylistics/home

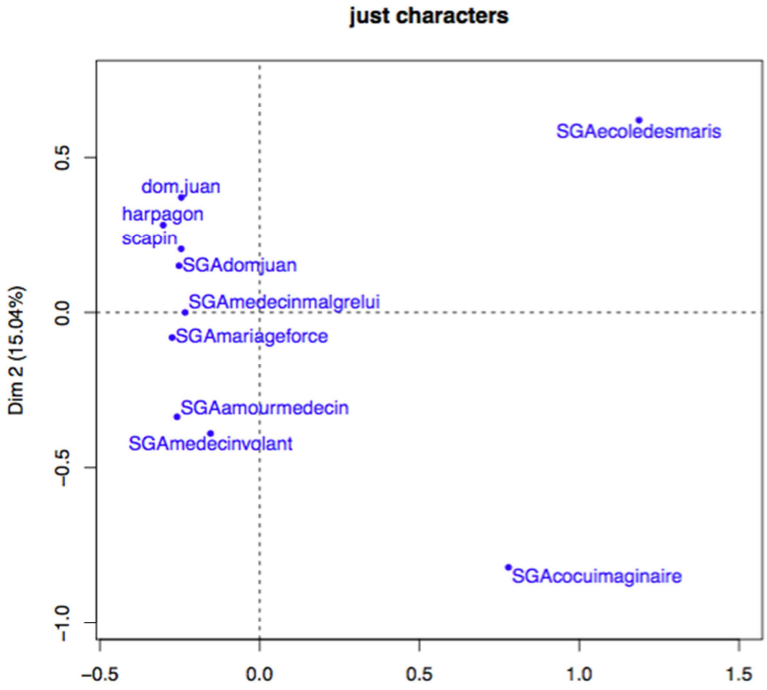

Figure 31. Correspondence analysis of four memorable Molière protagonists and the Sganarelles (A)

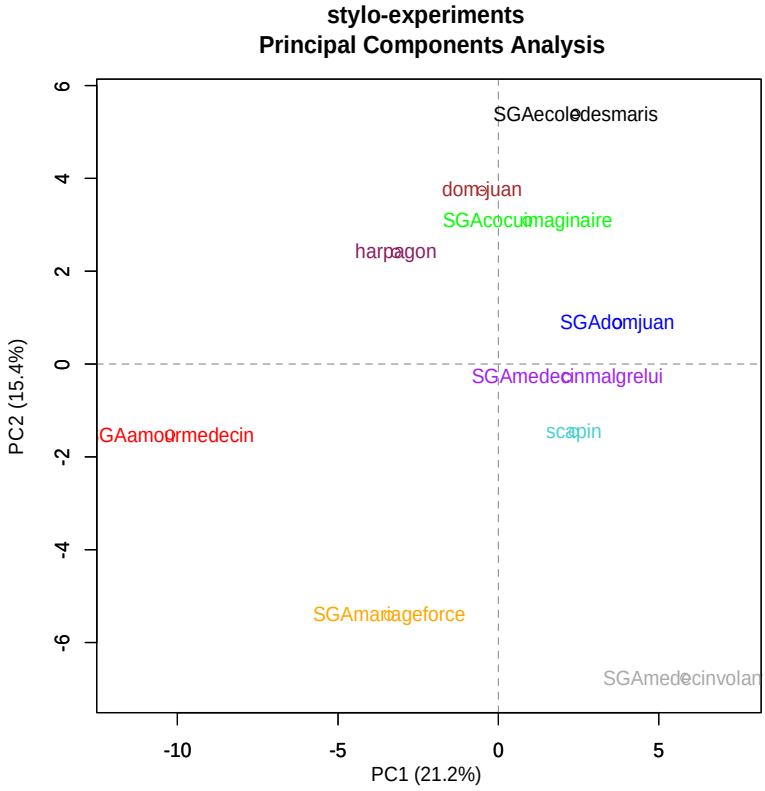

Figure 32. Lexical-based correspondence analysis of Molière's characters

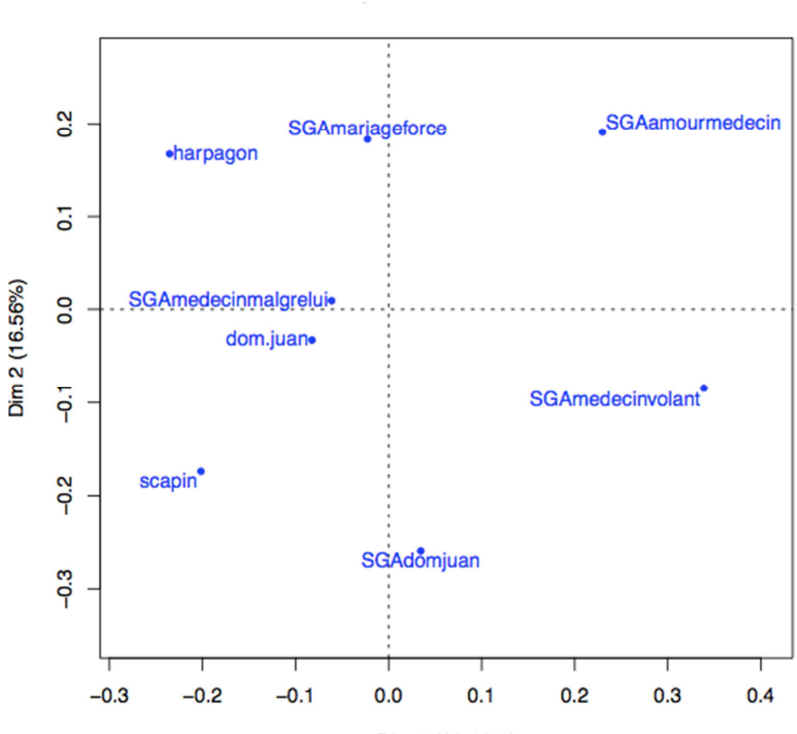

Figure 33. Correspondence analysis of four memorable Molière protagonists and the Sganarelles (B)

We clearly notice (in Figure 32) how the two Sganarelles in verse (that of *Ecole des maris* and that of *Cocu imaginaire*) are not isolated on one axes here. This is another indicator that lexicon is less sensitive to genre and register variation than syntax.

In order to carry out a more fine grained analysis we leave out the two Sganarelles in verse.

From the produced plots in Figure 33 and 34, we can notice that the Sganarelles does not clearly cluster in one homogenous group with respect to other characters. In fact Sganarelle of *Medecin malgé lui* for instance is more close to Dom Juan than to other Sganarelles. This suggests that Molière did not intend to give a syntactically distinguished discourse for the Sganarelle characters.

Concerning another research question, the hypothesis that we wanted to verify is the hypothesis by Forestier, that the Sganarelle of *Medecin Volant* is not a typical Sganarelle, but a different type of character, that of the clever servant, and that it might not have been called Sganarelle at all. From what we can notice from the correpondance analysis projection is that Sganarelle of *Medicin Volant* is quite isolated, from the others but so are the Sganarelle of *Dom Juan* and that of *Amour Medecin*.

However, by taking a closer look to the most contributive pattern of the projection, we found out that the Sganarelle of *Medecin Volant* is indeed quite marked by distinctive patterns, such as (here closest 5):

- `<(PUN) (NOM) (NAM)>`
- `<(NOM) (NAM) (PUN)>`
- `<(ADV) (KON) (PRO)>`

- `<(NOM) (NAM) (PUN)> (PRO)>`
- `<(NOM) (PUN) (NOM)>`

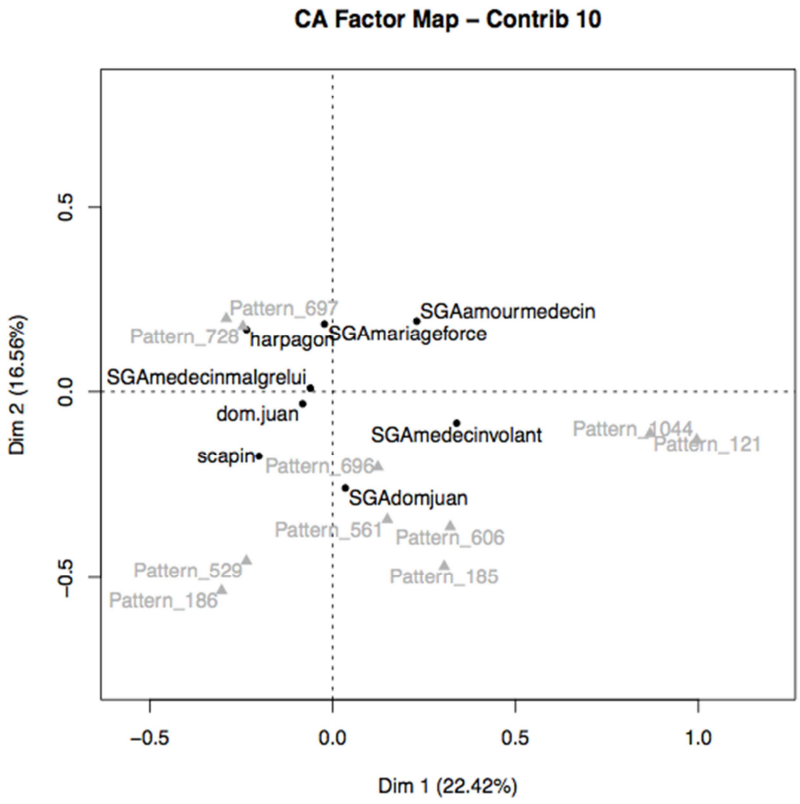

Figure 34. Correspondence analysis of four memorable Molière protagonists and the Sganarelles (C)

The instances of such patterns seem to show many repetitions; the discourse is not very complex and mostly addresses to some other character (notice the presence of many proper noun and noun tags, NAM and NOM. No verbs or adverbs appearing in the list). Actually, these results open the door up to more hypotheses. Is it a farcical character?

Also, notice how the Sganarelle of *Dom Juan* and Dom Juan himself aren't very similar. There seems to be no *play–fingerprint*. They aren't very similar even in terms of lexicon (see second plot in Figure 32); this is interesting, but should be further investigated.

Compare for instance with Sganarelle of Dom Juan. Here are his 5 most distinctive patterns:

- `<(NOM) (PUN) (PRO) (VER)>`
- `<(NOM) (PUN) (PRO)>`
- `<(PUN) (PRO) (VER) (ADV)>`
- `<(PUN) (PRO) (ADV) (VER)>`
- `<(NOM) (PRP) (NOM) (PUN)>`

The instances of such patterns seem to show a more complex discourse with relatively longer turns. This Sganarelle arguments, talks to himself, not only to the master and has articulate dialogues with other characters (notice the presence of verbs and adverbs tags, VER and ADV, in the list as opposed to the first one). As a conclusion, we can say there is no unquestionable and clear computational or visualization-based proof that Sganarelle of *Medecin Volant* is not prototypical with respect to the other ones, however the evidences that we gathered from such analysis suggest strongly that the fact of the matter is so.

5.4. Fourth Experiment: Molière's *"Raisonneurs"*

In our last experiment, we focus on the figure of the "*raisonneurs*", characters who take part in discussions with comical protagonists providing a counterpart to their follies. Such characters were interpreted at times as spokesmen for Molière himself, and the voice of reason, at other times as comical characters themselves and no less foolish than their opponents. Table 12 lists the plays we are going to analyze as well as the characters. Hawcroft's essay *Reasoning with fools* (2007) highlights the differences between five of these characters based on their role in the plot. Using this analysis as guidance, we compare significant linguistic patterns in order to see how these differences are marked by the stylistic choices of the author. Given the results of the previous experiment, we focus on the analysis of the discourse traits and on how they match to the communicative function each character needs to fulfill (Biber & Conrad 2009).

Table 12. Plays and characters

Play	Raisonneur	Counterpart
Ecole des femmes	Chrysalde	Arnolphe
Ecole des maris	Ariste	Sganarelle
Tartuffe	Cléante	Orgon
Misanthrope	Philinte	Alceste
Malade imaginaire	Béralde	Argan

Figure 35 shows the result of the correspondence analysis, with the five "*raisonneurs*" and the 10 patterns with the highest contribution labeled with their identifiers.

The relative distances between the characters seem to match what is already known from literary criticism; first of all Béralde, who is the only character to express himself in prose, is isolated on the right of the x-axis. As already remarked, it is not advisable to compare characters in prose and verse, but we have retained the example of Béralde to show how the proposed technique can easily identify differences in genre. As for the other characters, Hawcroft stresses the difference in the roles of Ariste, Philinte and Chysalde on the one hand and of Cléante on the other hand (something that is clearly reproduced in Figure 35). The latter is a more pro-active character,

more crucial to the plot; he is also less accommodating than the other three, who are depicted mostly as loyal friends and brothers, trying to help the hero to avoid the consequences of his foolish actions and beliefs. Instead, Cléante has also to worry about his sister's wellbeing: having to face not only the besotted brother in law, Orgon, but also the man who has dumped him, Tartuffe.

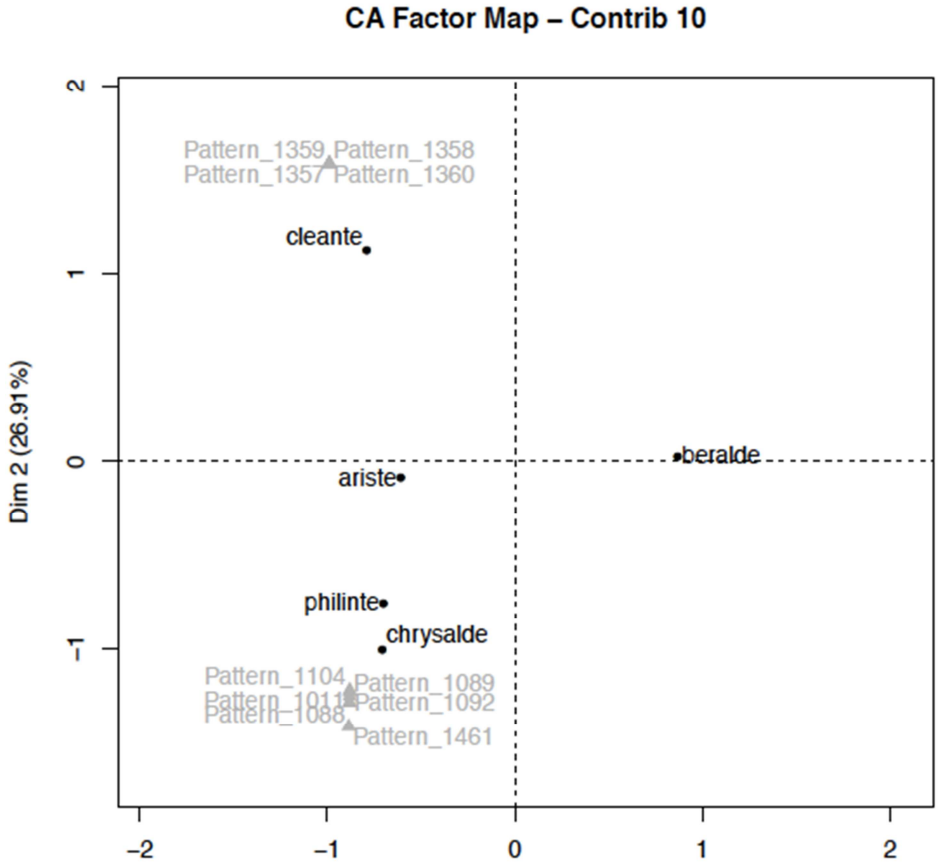

Figure 35. Top contributive patterns for the raisonneurs experiment

In order to confirm this intuition, it is necessary to turn our attention to what it is that exactly causes the spatial distribution, namely the high contribution patterns, we find above. As explained before, our technique allows us not only to find the corresponding pattern for each identifier on the plot, but also to extract all underlying instances in the texts. Some demonstrative analyses are performed.

Philinte and Chrysalde are strongly associated with patterns containing prepositional phrases separated by commas. Such patterns are used in contexts where the characters give advice in a very cautious, indirect way. The overuse of punctuation itself, in these two characters, seems to be an indication that the character should be played as a soft spoken person, who is fond of his friend and careful not to offend, e.g.:

Pattern (40): `<(,) (*) (PRP) (*) (NOM>`

Instances from **Chrysalde**:

- "Entre ces deux partis il en est un honnête **, Où dans l' occasion** l' homme prudent s' arrête ..."
- "Il faut jouer d' adresse **, et d' une âme** réduite , Corriger le hasard par la bonne conduite …"

Instances from **Philinte**:

- "**, Et pour l' amour** de vous , je voudrais , de bon cœur , Avoir trouvé tantôt votre sonnet meilleur ."

On the other hand, the patterns most associated with Cléante contain modal constructions, and are indicative of a more direct way of advising, and of stronger arguments, e.g.

Pattern (41): <(PRO:PER) (any word) (VER:infi) (PRP)>

- "Les bons et vrais dévots , qu' **on doit suivre à** la trace , Ne sont pas ceux aussi qui font tant de grimace ."
- "Et s' il **vous faut tomber dans** une extrémité , Péchez plutôt encore de cet autre côté ."

Finally, the patterns extracted for Béralde are indicative of the greater simplicity and repetitiveness of his prose, and of the stereotypical role he has in the play, which is that of a man concerned with his brother, as in:

Pattern (42): <(,) (DET:POS) (any word) (PUN)>

- "Oui **, mon frère ,** puisqu' il faut parler à cœur ouvert , …"

Just as for the experiment with the protagonists, this brief analysis is clearly meant not to provide new insights on the issue of *raisonneurs*[33] but rather to show that the system behaves in a consistent way with respect to basic and known assumptions on the plays. At the same time, it is possible to see how such an instrument can be used to investigate "old" issues from a new perspective, providing the researcher with new useful insights on Molière's use of language, and how syntactic structures and their underlying communicative functions can contribute to shape the linguistic profile of theatrical characters.

5.5. Discussion

This case study analysis shows us that some well-known traits of the analyzed protagonists can be automatically retrieved among the great mass of syntactic traits automatically extracted by sequential pattern mining.

[33] Hawcroft's analysis of the *raisonneurs* provided us with an interesting testing ground, but it should be compared to and read in the light of the influential interpretations of Georges Forestier on this topic.

At the same time, one important issue seems to emerge, concerning the relationship between communicative function and characterization when analyzing syntactic features. Indeed, it is clear that the two aspects cannot be fully disentangled. As has been demonstrated by recent developments in discourse analysis studies (Biber & Conrad 2009), the proportions of verbs to nouns, the use of pronouns and several other traits, differs according to register and to communicative situation. This explains the fact that the extracted patterns seem to be highly representative of the kind of situations in which the character finds himself in the plot, as well as of his/her station in life. Some distinguishing psychological traits (as in Harpagon) emerge, but are not as predominant as one might expect. This also tells us something particular about classical French plays where characterization was often left to the actor. Molière wrote most of his protagonists to be played by himself. So a lot of the characterization needs to be inferred by modern day performers rather than being explicitly given in the stage directions, and some room for freedom is left.

Nevertheless such studies seem interesting as they bring to the light the way in which the author constructed his characters and managed to give them each a voice that was plausible both from the social and the contextual point of view.

Chapter 6. Conclusion and Future Work

This chapter of the dissertation is intended to summarize the main contributions of our thesis, to highlight some of the open issues that still remain unsolved, and to suggest at the light of the produced results some directions for future work and further investigations.

6.1. Summary of the Contributions

In this dissertation, we focused on the extraction of complex yet computationally feasible stylistic features that are linguistically motivated, namely morpho-syntactic patterns, based on a hermeneutic unsupervised paradigm.

We have proposed a knowledge discovery process for stylistic characterization with an emphasis on the syntactic dimension of style by extracting relevant patterns from a given text without any prior knowledge. The proposed knowledge discovery process consists of two main steps, a sequential pattern mining pipeline followed by the application of some interestingness measures. In particular, the extraction of all possible syntactic patterns of a given length is proposed as a particularly useful way to extract interesting features in an exploratory scenario. Clearly the proliferation of patterns and the difficulty for humans to make sense of the huge amount of resulting dimensions of variation between texts is a major obstacle to this approach. We used interestingness measures in this scenario to treat and reduce such large quantities of dimensions. We evaluated and reported results on three proposed interestingness measures, each of which is based on a different theoretical linguistic background.

The experimental results indicate that the presented techniques are fairly effective in extracting interesting syntactic patterns, and this seems particularly promising as a computer-assisted literary analysis tool to support linguist and literary researchers in their critic analysis, especially if we take into account the unsupervised nature of this process.

From a comparative point of view, it is clear that the correspondence analysis-based interestingness measure was the best performing method both qualitatively and quantitatively. The strength of correspondence analysis lies in the fact that it allows users to easily identify the reasons why certain texts to be grouped together or to be scattered. This helps to overcome the lack of transparency in the presentation of results, something that often disappoints experts when faced with experiments using similar techniques. It is, therefore, well suited for syntagmatic approaches (such as the one we are using) that are per definition combinatorial and hence high-dimensional. Thus, the proposed methodology offers a useful instrument to facilitate literary analysis and

criticism; not only does it calculate and represent the distances between the analyzed texts, but it also provides a way to motivate and explain the differences based on the extraction of significant and distinctive sets of patterns for each character, which is a strong requirement for all computational stylistics methods. Moreover, we found out that the researchers from Labex OBVIl, with whom we have closely worked, have particularly appreciated the visualization aspect offered by this method compared to the other ones. Given its intrinsic nature, the distribution-peculiarity based measure seems very promising as well, especially for the analysis of such texts that, for their characteristics or for historical reasons, cannot support a comparative study as they are, in some way, unique. This might be the case of great poems from the antiquity, such as the *Iliad* or the *Odyssey* or even contemporary works whose style is too peculiar for comparison, such as James Joyces's *Ulysses*.

As part of our thesis contribution, we have taken the proposed methodology in our study case to a more specific application than the general stylistics purpose that is the study of stylistic characterization by analyzing the voice of Molière's characters in terms of distinguishing syntactic patterns.

Although the present dissertation focuses on classic French literary texts (novels and plays), the presented approach can be extended to any language and genre, provided a reliable automatic POS tagging for that language/genre is available.

It is also important to mention that our thesis is not meant to directly develop some sort of new stylistics theory, but rather to focus on the computational aspects that may help literature and stylistic researchers un examining their stylistic theoretical ideas (especially if it is related to syntax) and perhaps conceiving new ones.

In this line of thought, during our thesis we worked as well on the implementation of a computational stylistic tool concretizing the research work that we conducted during the thesis. This effort resulted in the development of a tool called EReMoS. The goal of EReMoS is to provide linguists and literature researchers with a computational stylistic tool conceived as a web application capable of extrac-ting and manipulating syntactic patterns through a simple, fast and ergonomic user interface.

Finally, even though it does not constitute an integral part of our core contribution detailed in this dissertation, it is also worth mentioning that we have conducted experiments and produced promising contributions with fairly accurate performances in the field of computational authorship attribution and authorship verification.

6.2. Open Issues and Future Work

Of course, the results of this thesis, although encouraging, are far from being perfect and show a number of limitations that affect our approach.

In fact, to begin with, the presented samples of syntactic patterns in the experimental evaluation chapter (Chapter 4) of this dissertation do not represent the overall quality of the extracted patterns. There exist patterns that do not show necessarily a relevant relation to any kind of identified stylistic preferences.

On the other hand, this methodology, as well as other similar ones, prompts the question of what is really captured by significant patterns. Some structures may be significant because they are

typical of an author's style, its fingerprint as we may say borrowing a metaphor often used in attribution studies, or they may be dictated by functional needs, due to the particular topic of the work, or to the conventions of the chosen genre. This is particularly true for syntactic analysis, where the functional constraints on the authorial freedom are more evident.

It is always hard in linguistics to separate the form from the function. For this reason, it is important to study syntactic patterns in the light of the sentences from which they are drawn (to avoid false conclusions). Nevertheless, the technique seems efficient in demoting those frequent constructions that are typical of French syntax in general without the need of a reference corpus; at the same time the syntactic structure of the extracted patterns and their use in vivid descriptions, in the presentation of characters and in the reconstruction of scenes (in the case study for instance) do seem to resonate with the particular use of language typical of the analyzed texts.

However, our work can be improved in several ways. In fact, in the light of the produced results and analysis, it is possible to outline some directions for future research work and investigations. We organize them in two different aspects: the literary and linguistic aspect on the one hand, and the technical aspect on the other hand.

Firstly, on the literary and linguistic aspect and using Biber & Conrad (2009)'s definitions, it is worth asking how far it is possible to distinguish style from register and genre when analyzing syntactic structure, especially considering that stylometric studies have traditionally focused on features such as word and sentence length, or lexical richness. In particular our methodology, based on the statistical properties of the extracted patterns, can invariably capture local changes in register motivated by the different elements of the novel (introduction, descriptions, scenes of action, dialogue, etc.), along with stylistic traits. Thus, we think that this point should be deeply investigated. Producing an expert manual extraction and annotation of the patterns on a test set, serving as a basis for a more robust quantitative and qualitative evaluation, can be a possible way to handle this issue.

Moreover, the analysis of syntactic choices of the theatrical prose, as emerging from the combination of contiguous syntactic categories, can provide us with a different and interesting insight in texts even of a relatively short length such as the ones analyzed in our work. In particular, morpho-syntactic patterns could also be used to compare the features of theatrical dialogue to those of genuine spoken dialogues, in order to investigate how far theatrical prose is able to mimic speech and real oral interaction. Another interesting area of research that may benefit from the proposed approach beyond the study of characters is the investigation of Molière's dialogues on a more typological level, comparing for instance different types of scenes (long monologues, the comic exchanges, etc.) as done by Gabiel Conesa (1983) in his study. Here too, already known distinctive features could be provided with additional supporting corpus evidence, by the bottom up extraction and filtering of distinctive syntactic patterns.

Secondly, on the technical aspect, the most obvious issue that one can point out is the accuracy of the syntactic POS tagger used as part of the syntactic analysis processing chain. It is well known that the majority of the available POS taggers are trained on journalistic data. Thus, their generalization performance on the literary texts, which have their own specific linguistic properties, is quite far from perfect. Even though the produced annotation could be fairly acceptable for prose texts, it sometimes extremely bad, yet understandable, how such POS taggers fail to recognize the syntactic structures of poetic texts for instance. This constitutes a barrier for the application of syntax–based computational stylistics work on such texts. The solution that we propose to solve this issue is to work on a domain-specific POS tagger able to handle the peculiarity of literary texts. This can be achieved through a domain adaptation process by retrai-

ning the POS tagger on a sufficient set of manually-corrected annotations performed on literary texts for instance.

On another technical perspective, the analysis carried out so far is somehow static in the sense that each pattern is analyzed separately from others, without taking into account the overall dynamic in which those patterns find themselves and the relationships that may exist between them. Another related point is that each text is investigated ignoring the time it was written in, consequently the stylistic evolution of someone's' writings is not taken into consideration. To deal with this issue, we could refer to some techniques from the knowledge extraction domain. More specifically, a suitable solution could be to extract linguistic summaries in the form of gradual patterns capable of expressing relations of correlation and co-variation between different entities and highlighting the dynamic between them, for instance between the different texts of a corpus, texts with respect to time, syntactic patterns or categories with respects to each other, and so on.

Appendix A. EReMoS: A Computational Stylistics Tool for Extracting and Searching Syntactic Patterns

As part of our contribution, we worked on the implementation of a computational stylistic tool partly concretizing the research work that we have conducted during the thesis. This effort resulted in the development of a tool called EReMoS ("**E**xtraction et **Re**cherche de **Mo**tifs **S**yntaxiques").

Basically, EReMoS is a computer-assisted computational stylistic tool conceived as a web application and developed in the ACASA team at the computer science laboratory of Paris 6 (LIP6). This web application allows extracting and searching syntactic patterns from/in an uploaded text. It can be accessible at this address:

http://eremos.lip6.fr/

The goal of this tool is to be capable of manipulating syntactic patterns in order to assist linguists and literature researchers in their studies on syntactic styles. These studies may include the recognition of syntactic structures or features characterizing a special type of texts such as texts written by a certain author or texts belonging to a specific literary genre.

Our main aim is to provide an easy and intuitive access to the features implemented by the tool through a simple, fast and ergonomic user interface. Indeed, this tool is suitable for linguists or literary researchers that want to conduct some computational and quantitative analysis of the syntactic aspect of some texts without having the necessary computer skills to manipulate natural language processing tools such as the POS taggers.

Technically speaking, EReMoS consists of two main components:

- An extraction engine that is responsible of analyzing the syntactical structure of the uploaded text (text under investigation), transforming the results of this syntactic analysis into a numerical format that could be algorithmically handled, and then mining these results for the extraction of syntactic patterns, along with other useful quantitative information, all of which is done on the basis of some user-specified parameters.
- A search engine that is responsible of searching and identifying the textual instances of some syntactic patterns in the analyzed text and thus adding the lexical information to the syntactic one.

The extraction engine of EReMoS is built upon SPMF[34] which is an open-source data mining library written in Java specialized in sequential data mining. In this component, we make use of TreeTagger[35] for the syntactic analysis part as well.

At first, EReMoS was developed to analyze texts written in Franch language. It was then extended to support German texts as requested by researchers from the GCDH:: Göttingen Dialog in Digital Humanities[36] following its presentation in the Göttingen Dialog in Digital Humanities in 2015. We are planning to extend EReMoS to other languages in the future.

Before going into the details of the features that EReMoS offers, it is worth mentioning that many computer-assisted literary tools have been developed, especially in the francophone community. We do not intend to make an exhaustive listing of all those tools in this appendix but as an illustration we can mention: TXM[37], Lexico[38] and Hyperbase[39] for instances. What those tools have in common is that they focus mostly on the lexical part of the texts/corpus which is in fact easier to count and analyze. However, they incorporate a very useful and interesting bench of features and analysis methods including statistical analysis methods (correspondence analysis–like methods for instance). Somehow, this advantage can turns into a weakness in the sense that the tool could be judged as complex by users who are unfamiliars with such methods. Technically speaking, most of those tools are developed as desktop application. Thus, it is necessary for someone who wants to use them and by the way beneficiate from their analysis capabilities, to install them locally in a computer. This could also be considered as a questionable point.

In what follows, we give an overview of the implemented features:

Extracting patterns

The text chosen by the user will be imported from their personal files. To perform an extraction of the syntactic patterns corresponding to its needs, the user should specify some parameters as illustrated in Figure 36:

- The minimum and maximum pattern's size
- The minimum relative threshold of pattern's support with respect to the text size
- The minimum absolute threshold of pattern's support

Presenting the results

Once the extraction performed, the user can observe the extracted information. Indeed, the tool provides him, as illustrated in Figure 37, with a pie-chart representing the proportions of the syntactic POS tags present in the text, and a table presenting the extracted syntactic patterns alongside their supports.

[34] http://www.philippe-fournier-viger.com/spmf/

[35] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[36] http://www.gcdh.de/en/

[37] http://textometrie.ens-lyon.fr/spip.php?rubrique64

[38] http://www.tal.univ-paris3.fr/lexico/lexico3.htm

[39] http://ancilla.unice.fr/

Exploring the extracted patterns

Filtering options may be added in order to limit the result's size according to the interest of the user. To perform such operation, we have implemented a regex-like filtering process.

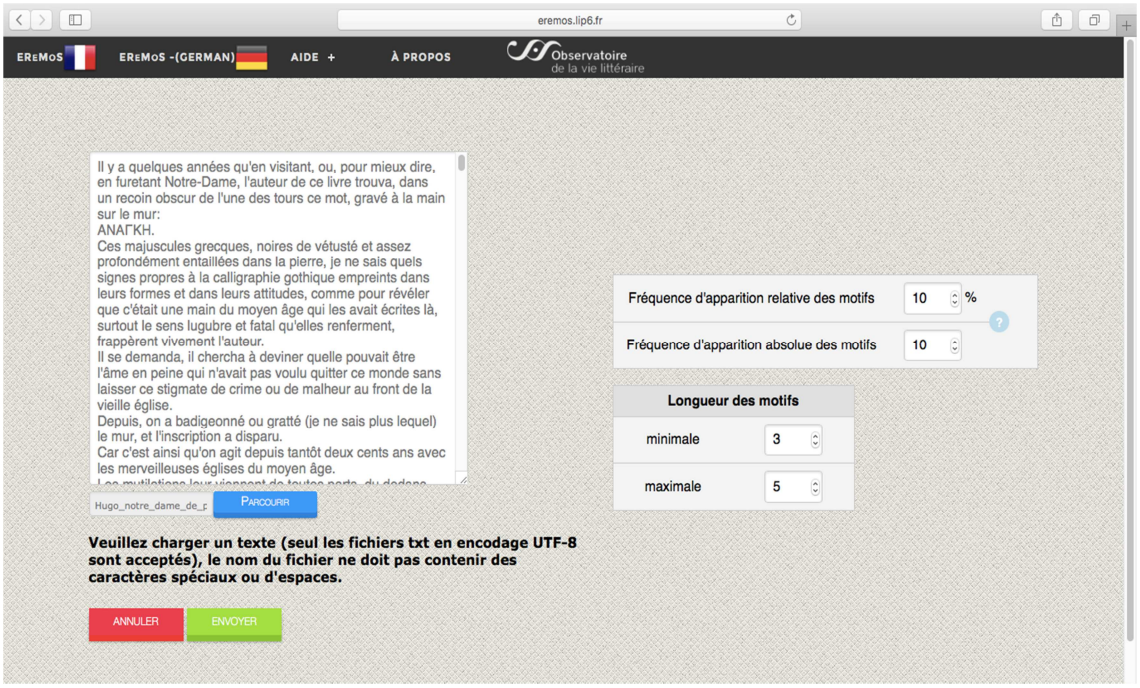

Figure 36. EReMoS homepage in which the user uploads the text and specifies the extraction parametrs

Searching the instances of some pattern

As illustrated in Figure 38, the user can search and explore in the text the textual instances of a selected pattern. These instances will be highlighted accordingly in the text. The user has the possibility to visualize the positions' distribution of some pattern using a bar plot (see Figure 39) as done in the distribution-peculiarity interestingness measure (see Figure 24 and go back to Subsection 3.3.2.3 of Chapter 3 to have more information about this visualization).

Exporting the results for an offline exploration

Finally, the user has the possibility to export the produced results, both the extracted patterns and the textual instances, in an adapted format for an offline exploration.

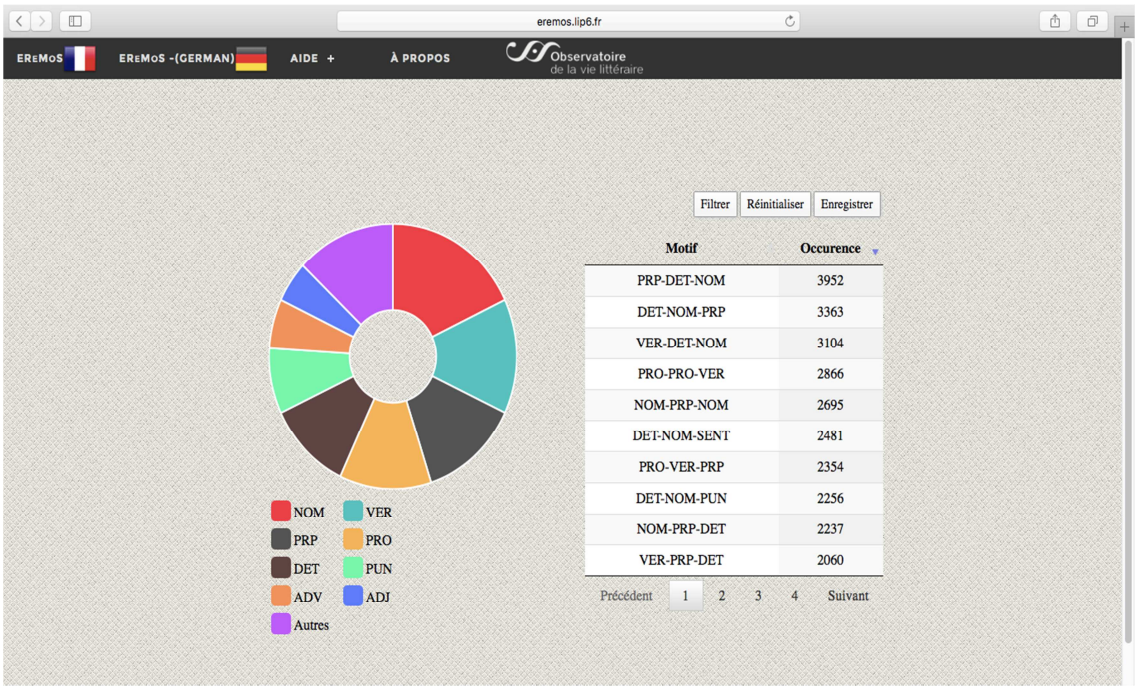

Figure 37. The Results page containing the POS tags' proportions and the extracted patterns

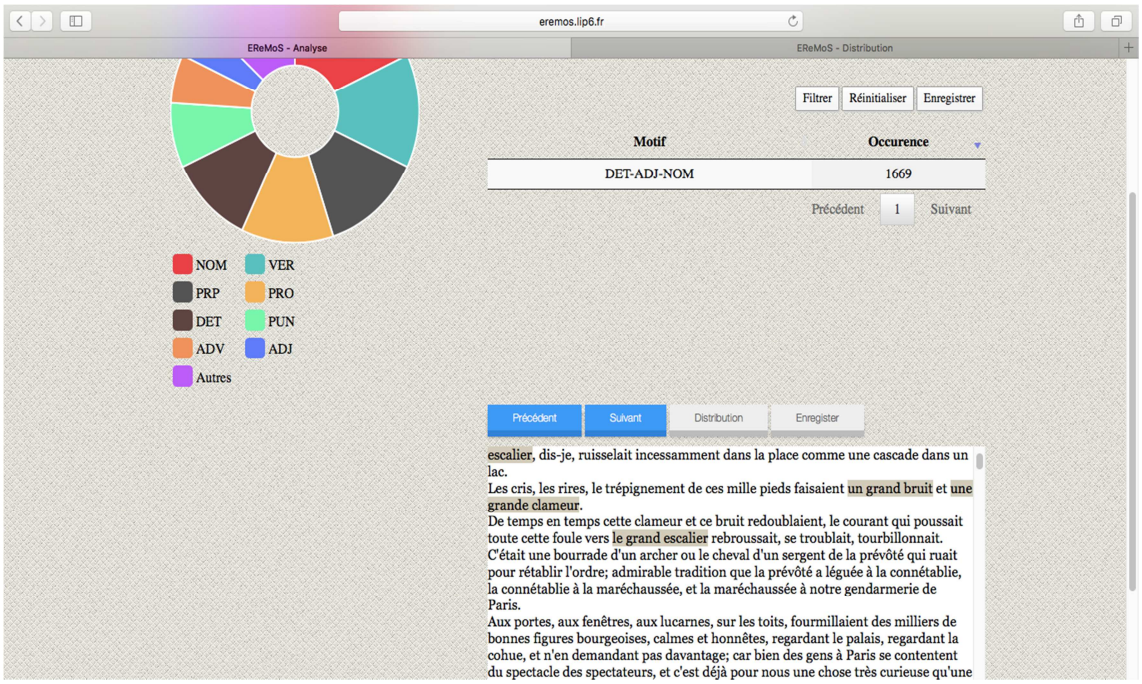

Figure 38. Exploring the pattern's textual instances in EReMoS

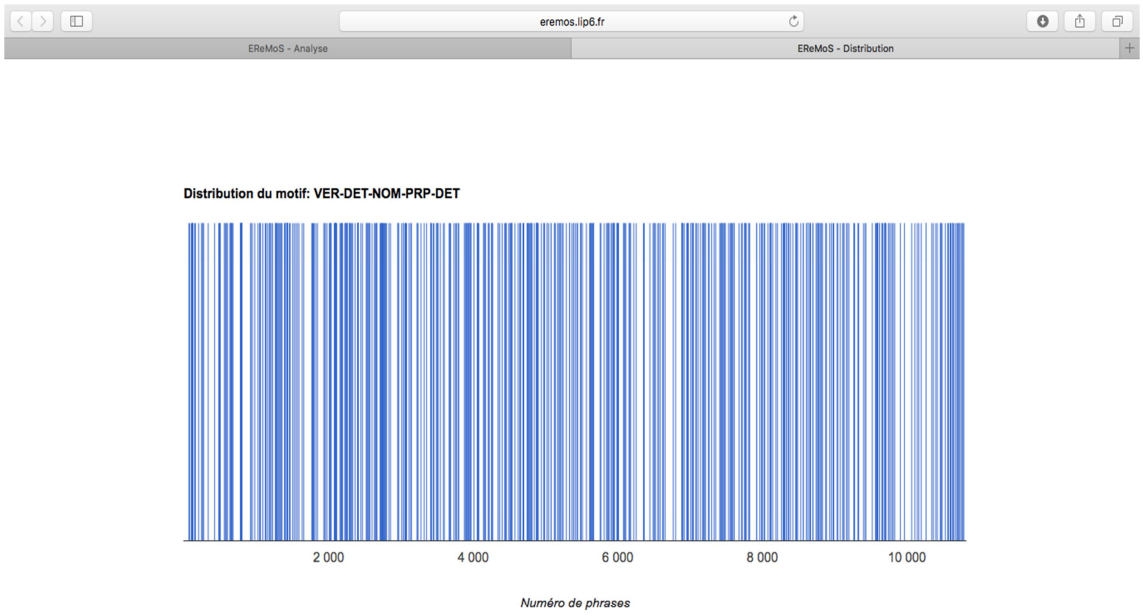

Figure 39. The bar plot illustrating the pattern's distribution in the text

Appendix B. Evaluating the Effectiveness of Sequential Rule-based Features for Authorship Attribution

Many stylistic features have been used as style markers for the task of authorship attribution[40]. Function words are shown to be a very reliable and effective indicator of authorship, so they are suited to handle the task of authorship attribution or some other related tasks such as authorship verification. In fact, function words have little lexical role to play, but instead they serve mostly a syntactic role by expressing grammatical relationships among words or collections of words within a sentence.

In an effort to develop more complex yet computationally feasible stylistic features that are more linguistically motivated, Hoover (2003) pointed out that exploiting the sequential information existing in the text could be a promising line of work. He proved that frequent word sequences and collocations can be used with high reliability for stylistic attribution. In this line of research, here we study the problem of authorship attribution in classic French literature. Our aim is to evaluate the effectiveness of style markers extracted using sequential data mining techniques for authorship attribution. In this contribution, we focus on extracting style markers using sequential rule mining. We compare results given by these new style markers to that of the state-of-the-art features like function words frequencies and POS-tag n-grams, and we assess whether this type of markers is sufficient for accurate identification of authors. In what follow, we briefly explore this point of using sequential rules as features in Section B.1. Then, we present the experimental set up in Section B.2 in which we first describe the data set used in the experiment, and then the

[40] Go back to Section 2.5 of Chapter 2 for more details

classification scheme and algorithm employed for this experiment. The results and discussions are presented at the end of this appendix in Section B.3.

B.1. Sequential Rules as Stylistic Features

At this point, as an illustration of the idea prompted above, we propose to explore the predictive property of stylistic features based on sequential rule mining. So as main experiment in our work, we studies the stylistic characterization of 10 classic French authors using different stylistic features ranging from a relatively low linguistic level to a more high and complex one. We chose to focus on the syntactic aspect of style, so as stylistic features in this experiment we took:

- Frequency of function words
- Sequential rules of function words
- Tri-gram of POS tags
- Sequential rules of POS tags

From the list below, the frequencies of functions words are obviously the least complex linguistics features and subsequently the least relevant and interesting one to characterize the style of an author. They neither serve explicit stylistic lexical preferences, nor an explicit stylistic syntactic trait. The other stylistics features are linguistically more complex and stylistically more interesting. For instance the sequential rules of functions words can capture the differences between the periodic and the loose styles. While, the sequential rules of POS tags can play an alternative role to the grammatical productions rules used in formal grammar (O'Neill & Ryan 2001) except that in this case, those rules will gives insights about the syntactic choices of an author rather than describing the grammar in a general way as done using the productions rule.

What one should expect from such configuration is that the more relevant the feature is to describe the stylistics choices of a given author, the more able and suitable it is to distinguish his writing from that of different author. That is to say, for a stylistic characterization based on classifcation approach, one would expect the sequential rules of function words to be more effective that the frequencies of function words since they are more stylistically relevant (they are able to tell us more about the writing style of an author, and they are more easy to interpret in the same time as well). Similarly, we would expect sequential rules of POS tags to be more effective of both of them for the same reason.

Our aim in this experiment is to test the validity of this hypothesis by evaluation the effectiveness of stylistics features presented above in the context of authorship attribution. Well, it turn out that this hypothesis is not true, at least for the corpus that we have considered in this experiment. This can be considered as a clear argument suggesting that the less complex features, acting on a relatively low linguistic level, are more suitable for the authorship studies from a classification point of view.

B.2. Experimental Settings

B.2.1. Data Set

To test the effectiveness of sequential rules with respect to POS-tag and function words for authorship attribution, we use texts written by: Balzac, Dumas, France, Gautier, Hugo, Maupassant, Proust, Sand, Sue and Zola. This choice was motivated again by our special interest in studying the classic French literature of the 19th century. Our choice of authors was also affected by the fact that we want to cover the most important writing styles and trends from this period. For each of the ten authors mentioned above, we collected 4 novels, so that the total number of novels is 40. The next step was to divide these novels into smaller pieces of texts in order to have enough data instances to train the attribution algorithm. Researchers working on authorship studies on literature data have been using different dividing strategies. For example, Hoover (2003) decided to take just the first 10,000 words of each novel as a single text, while Argamon & Levitan (2005) treated each chapter of each book as a separate text. In our experiment, we chose to slice novels by the size of the smallest one in the collection in terms of number of sentences; more information about the data set used in the experiment is presented in Table 13.

Table 13. Statistics for the data set used in our experiment

Author Name	# of words	# of texts
Balzac, Honoré de	548778	20
Dumas, Alexandre	320263	26
France, Anatole	218499	21
Gautier, Théophile	325849	19
Hugo, Victor	584502	39
Maupassant, Guy de	186598	20
Proust, Marcel	700748	38
Sand, George	560365	51
Sue, Eugène	1076843	60
Zola, Émile	581613	67

B.2.2. Classification Scheme

In the current approach, each text was segmented into a set of sentences (sequences) based on splitting done using the punctuation marks of the set {'.','!','?',':','...'}, then the corpus was POS tagged and function words were extracted. The algorithm described in (Fournier-Viger and Tseng, 2011) was then used to extract sequential and association rules over the function words and the POS-tag sequences from each text. These rules will help us gather not only sequential information from the data, but also structural information, due to the fact that a text characterized by long sentences will result in more frequencies of the rules.

Each text is then represented as a vector R_K of frequencies of occurrence of rules, such that $R_K = \{r_1, r_2, \dots, r_K\}$ is the ordered set by decreasing normalized frequency of occurrence of the top K rules in terms of support in the training set. Each text is also represented by a vector of

normalized frequencies of occurrence of function words and POS-tag 3-grams. The normalization of the vector of frequency representing a given text was done by the size of the text.
Our aim is first to compare the classification performance of the top K function word-sequential rules (SR) (Top 100, 200, 300 SR were examined) to the function words frequencies. And secondly to compare top K POS-tag sequential rules to the POS-tag 3-gram frequencies (Top 100, 200, 300, 400, 500, 600, 700, 800 SR were examined).

Given the classification scheme described above, we used SVMs classifier to derive a discriminative linear model from our data. To get a reasonable estimation of the expected generalization performance, we used 5-fold cross–validation. The dataset was split into 5 equal subsets; the classification was done 5 times by taking 4 subsets for training in each time and leaving out the last one for testing. The overall classification performance is taken as the average performance over these 5 runs. For evaluating the attribution performance, we used the common measures used to evaluate supervised classification performance: we have calculated precision (P), recall (R), and F-measure (F_1).

B.3. Results and Discussion

Results of measuring the attribution performance for the different feature sets presented in our experiment setup are summarized in Table 14 for features derived from function words, and in Table 15 for those derived from POS-tag. These results show in general a better performance when using function words, which achieved a nearly perfect attribution (e.g., F_1 =0.99 for FW frequencies and F_1 = 0.939 for Top 300 FW-SR), over POS-tag features.

Table 14. 5-fold cross-validation results for our data set. SR refers to Sequential Rules. FW refers to Functions words

Feature set	**P**	**R**	**F_1**
Top 100 FW-SR	0.901	0.886	0.893
Top 200 FW-SR	0.942	0.933	0.937
Top 300 FW-SR	0.940	0.939	0.939
FW frequencies	**0.990**	**0.988**	**0.988**

But contrary to our hypothesis, function word frequency features, which fall under the bag-of-word assumption known to be blind to sequential information, outperform features extracted using sequential rule mining technique. The same thing can be said for the POS-tag 3-grams.

By analyzing the individual attribution performance for each author separately, we notice a significant variance between the attribution performance of one author and that of another (e.g., F_1 = 1 for Proust comparing to F_1 = 0.673 for Dumas), some individual results are presented in Table 16. This particularity is due to the fact that some authors have more characterizing style than others in their works used for the experiment.

Table 15. 5-fold cross-validation results for our data set. SR refers to Sequential Rules. POS refers to Part-Of-Speech

Feature set	**P**	**R**	**F_1**
Top 100 POS-SR	0.743	0.725	0.733
Top 200 POS-SR	0.728	0.703	0.715
Top 300 POS-SR	0.831	0.817	0.823
Top 400 POS-SR	0.847	0.833	0.839
Top 500 POS-SR	0.859	0.841	0.849
Top 600 POS-SR	0.87	0.855	0.862
Top 700 POS-SR	0.885	0.869	0.876
Top 800 POS-SR	0.886	0.875	0.880
POS 3-gram frequencies	**0.991**	**0.990**	**0.990**

Table 16. Individual 5-fold cross-validation results for each author evaluated for the Top 700 POS-tag sequential rules

Author Name	**P**	**R**	**F_1**
Balzac	0.880	0.750	0.809
Dumas	0.655	0.693	0.673
France	0.920	0.960	0.939
Gautier	0.950	0.850	0.897
Hugo	0.887	0.950	0.917
Maupassant	1.00	0.85	0.918
Proust	1.00	1.00	1.00
Sand	0.925	0.901	0.912
Sue	0.861	0.866	0.863
Zola	0.985	1.00	0.992

Actually despite the fact that they are not much relevant features to describe the stylistic characterization, there is an agreement among different researchers that function words are the most reliable indicator of authorship. There are two main reasons for this property. First, because of their high frequency in a written text, function words are very difficult to consciously and voluntarily control, which makes them more inherent trait and consequently minimizes the risk of false attribution. The second is that function words, unlike content words, are more independent from the topic or the genre of the text, so one should not expect to find great differences of frequencies across different texts written by the same authors on different topics (Chung and Pennebaker, 2007). Yet, they are basically relying on the bag-of-words assumption, which stipulates that text is a set of independent words. This assumption completely ignores the fact that there is a syntactic structure and latent sequential information in the text. De Roeck (2004).

As we have seen, it turn out the hypothesis pointed in the beginning of this section is not true, at least for the corpus that we have considered in this experiment. This can be considered as a clear argument suggesting that classification approaches are not that suitable for the stylistic characterization studies. In fact, there is a difference between the characterizing ability of a stylistic feature in one hand and its discriminant power in the other hand. The most relevant and suitable stylistic features to perform a discriminant task such as stylistic classification are the one that operate on the a low linguistic levels such as function words, and that are subsequently more difficult to linguistically interpret and understand and does not necessarily enhance the knowledge concerning the style of the text from which they were extracted.

Appendix C. Anomaly Detection Approach for Authorship Verification

Authorship verification is a special case of the authorship attribution problem. In the authorship verification problem though, we are given samples of texts written by a single author and we are asked to assess if a given different text is written by this author or not (Koppel et al. 2009). As a categorization problem, modifying the original attribution problem in this way makes the task of authorship verification significantly more difficult partly because building a characterizing model of one author is much harder than building a distinguishing model between two authors (Koppel & Schler 2004).

Authorship verification has two key steps. First, an indexing step based on style markers is performed on the text using some natural language processing techniques such as tagging, parsing, and morphological analysis. Then, an identification step is applied using the indexed markers to verify the validity of the authorship. The verification step can be addressed as a one-class problem (written-by-the-author) or as a binary classification problem (written-by-the-author as positive vs. not-written-by-the-author as negative). However, both of these formulations of the problem have drawbacks. In the case of binary classification, one should collect a reasonable amount of representative texts of the entire "not-written-by-the-author" class, which is difficult, if not impossible. In

the case of one-class classification, one does not take advantage from negative examples that we do not actually lack for them even though they are not representative of the entire class.

In this appendix part of the thesis, we address the authorship verification problem as an anomaly detection problem where texts written by the candidate author are seen as normal data while texts not written by that author are seen anomalous data. We propose an anomaly detection approach with two different variations: the first variation is based on a weakly-supervised probabilistic model and the second variation is based on an unsupervised distance-based model. However, both of them are based on a multivariate Gaussian distribution.

The rest of this appendix part is organized as follow. We first give an overview of the anomaly detection problem in Section C.1 and then describe our approach in Section C.2. We discuss the relevance of the chosen stylistic features in Section C.3. We then experimentally validate the proposed method in Section C.4 and C.5 using a classic French corpus. Finally, we use the best performing method, which is the weakly supervised probabilistic method, to settle a literary mystery case in Section C.6.

C.1. Anomaly Detection

Anomaly detection is a challenging task which consists of identifying patterns in data that do not conform to expected (normal) behavior. These non-conforming patterns are called anomalies or outliers (Chandola et al. 2009). Anomaly detection has been successfully used in many applications such as fault detection, radar target detection and hand written digit recognition (Markou & Singh 2003).
This technique has also been used to deal with textual data for various purposes such as detecting novel topics, events, or news stories in a collection of documents or news articles (Chandola et al. 2009). Anomaly detection is based on the idea that one can never train a classification algorithm on all the possible classes that the system is likely to encounter in real application. Anomaly detection is also suitable for situations in which the class imbalance problem can affect the accuracy of classification (see Figure 40) (Wressnegger et al. 2013).

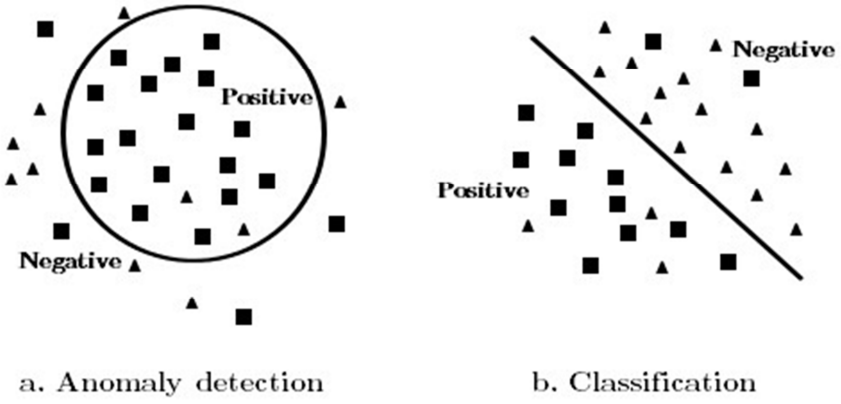

Figure 41. The anomaly detection and the classification learning schemas

Many anomaly detection techniques fall under the statistical approach of modeling data based on its statistical properties and using this information to estimate whether a test sample comes from the same distribution or not (Markou & Singh 2003). Another common method for anomaly detection is the one-class SVM that determines a hyper sphere enclosing the normal data (Heller et al. 2003). In what follows, we describe and use two anomaly detection methods for authorship verification that straightforwardly follow the definition given above. These two methods are discussed in the next section.

C.2. Proposed Approach

In our approach, we address the authorship verification task as an anomaly detection problem where texts written by a given author X are seen as normal data, while texts not written by that author X are seen anomalous data. In this section, we describe the first variation based on an unsupervised distance-based model, then the second variation based on the weakly supervised probabilistic model.

C.2.1. Unsupervised Distance-based Medel

The approach to anomalous text detection is to train a n-dimensional multivariate Gaussian model on the style markers extracted from sample of text written by an author X. Every newly arriving text (data instance) that we went to verify as written by X or not is contrasted with the model of normality, and a distance is computed. In fact, for n-dimensional multivariate normally distributed data; the values are approximately chi-square distributed with n degrees of freedom. In this case, the multivariate outliers can simply be defined as observations having a large squared Mahalanobis distance (Filzmoser 2004). Thus, the computed distance describes the likelihood of the new text to have been written by X compared to the average data instances seen during the training. If the distance surpasses a predefined threshold $\propto$, the instance is considered an anomaly and the text is considered not being written by the author X.
As a threshold, the quantile of the chi-square distributed (eg., 97,5% quantile) can be considered. Such method has been already successfully used (Rousseeuw & Van Zomeren 1990).

The method can be formulated into three steps as follow: Let $x^{(i)}$ be a n-dimensional vector (A vector of n style markers' frequency) representing the i-text in the training data set containing m texts. $(1 \leq i \leq m)$:

1. Train a Multivariate Gaussian distribution model M on the normal data. This is done by estimating the two distribution parameters: the multivariate location μ and the covariance matrix Σ:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \qquad (1)$$

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^{T} \qquad (2)$$

2. Given a new instance x, compute the Mahalanobis distance $\mathrm{D_M}(x)$:

$$\mathrm{D_M}(x) = \sqrt{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)} \qquad (3)$$

3. Predict the anomaly ($y = 1$) of the instance x given the distance threshold $\propto$:

$$y = \begin{cases} 0 & \text{if } D_M(x) < \propto \\ 1 & \text{if } D_M(x) \geq \propto \end{cases} \quad (4)$$

For the experimentation, two different thresholds have been considered: $\propto_1$ for the 95% Chi-squared quantile and $\propto_2$ for the 97,5% Chi-squared quantile.

C.2.2. Weakly Supervised Probabilistic Model

Unlike the first method, in this second variation we use a probabilistic anomaly detection method that can benefit from anomalous examples for the authorship verification process which is also based on a multivariate Gaussian modeling. Given the fact that unsupervised anomaly detection approaches have difficulties to match the required detection rates in many tasks and there exists a need for labeled data to guide the model generation (Görnitz et al. 2014), this method is weakly supervised in the sense that it takes into consideration a small amount of representative anomalous data for the model generation.

The approach to anomalous text detection is the same as the previous method for the first step. That is, on have to train a multivariate Gaussian distribution model on the style markers extracted from sample of text written by an author X. Then, every newly arriving text (data instance) that we went to verify as written by X or not is contrasted with the probabilistic model of normality, and in this case, a probability of normality is computed instead of a distance. The probability describes the likelihood of the new text to have been written by X compared to the average data instances seen during the training. If the probability does not surpass a predefined threshold λ, the instance is considered an anomaly and the text is considered not to have been written by the author X. To define the probability threshold, we cross-validate over a data set containing both anomalous and non-anomalous data and we set the threshold to the value that maximizes the authorship verification performance on this cross-validation data set. This threshold is used afterward on the test set.

As before, the method can be formulated into three steps as follow: Let $x^{(i)}$ be a n-dimensional vector (A vector of n style markers' frequency) representing the i-text in the training data set containing m texts. $(1 \leq i \leq m)$:

1. Train a Multivariate Gaussian distribution model M on the normal data. This is done by estimating the two distribution parameters: the multivariate location μ and the covariance matrix Σ:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \quad (5)$$

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^T \quad (6)$$

2. Given a new instance x, compute the probability $p(x)$:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) \quad (7)$$

3. Predict the anomaly ($y = 1$) of the instance x given the probability threshold λ:

$$y = \begin{cases} 1 \; if \; p(x) < \lambda \\ 0 \; if \; p(x) \geq \lambda \end{cases} \qquad (8)$$

C.3. Considered Style Markers

The nature of the style markers used as attributes to describe and to get the n-dimensional vector representing the text is very important and determines the applicability of our method. In fact, the nature of these attributes should respect the Gaussian assumption made to train the multivariate Gaussian model.

Table 17. List of the French function words used in our experiment

1. le	11. des	21. en
2. la	12. du	22. qui
3. l'	13. d'	23. elle
4. un	14. je	24. dans
5. une	15. au	25. qu'
6. sa	16. de	26. pour
7. s'	17. et	27. vous
8. son	18. à	28. plus
9. ce	19. il	29. sur
10. les	20. que	30. on

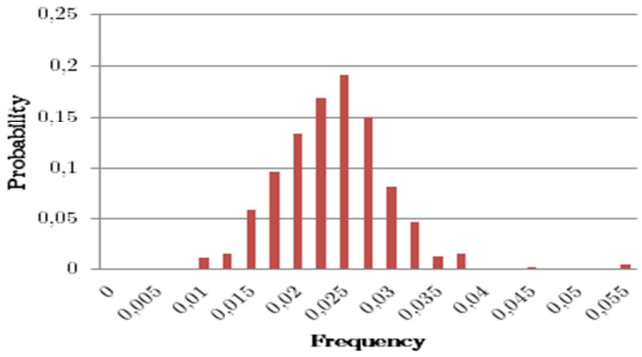

Figure 42. The probability of frequency of the French function word “de” has a Gaussian behavior

For our experiment, we chose to test these methods on two types of style markers separately. Each text in our data set is mapped onto a vector of the frequency of the most frequent function words (see Table 17) and a vector of the frequency of POS-tags.
There are two main reasons for using the frequency of function words as attributes. First, because of their high frequency in a written text, function words are very likely to have a Gaussian behavior (see Figure 41). Secondary, function words, unlike content words, are difficult to consci-

ously control, thus they are more independent from the topic or the genre of the text (Chung & Pennebaker 2007). In fact, Koppel & Schler (2004) found that all the work of distinguishing the styles of different authors is accomplished with a small set of features containing frequent function words. Based on that information and to get a right balance between the features-set size and the dataset size, we limit our study to the most 30th frequent function words. The part-of-speech-based markers are also shown to be very effective because they partly share the advantages of function words.

C.4. Experimental Settings

C.4.1. Data Set

To test the effectiveness of our method, we used novels written by: Balzac, Dumas and France. This choice is done as to get a challenging problem, since these three authors are knows to have relatively comparable syntactic styles. More information about the data set used for the experimentation is summarized in Table 18. For each of the three authors mentioned above, we collected 4 novels, so that the total number of novels was 12. The next step was to divide these novels into smaller pieces of texts in order to have enough data instances (artificial documents) to train and test the probabilistic model.

Table 18. Data set used in our experiment

Author Name	# of texts
Balzac, Honoré de	126
Dumas, Alexandre	190
France, Anatole	128

In this experiment, we chose to chunk each novel into approximately equal parts of 2000 words, which is below the threshold proposed by Eder (2013) specifying the smallest reasonable text size to achieve good attribution. This increases the degree of the difficulty of the task.

C.4.2. Verification Protocol

In the experiment of the unsupervised distance-based method, function words were first extracted. Each text is then represented by a vector $R_{30} = \{r_1, r_2, \dots, r_{30}\}$ of normalized frequencies of occurrence of the top 30 function words in the corpus. Then, for each author, we used 75% of the data generated by texts written by him to estimate the parameters of the representing multivariate Gaussian model, and 25% of the data from each author as testing set.
In the experiment of the weakly supervised method, we consider both the two types of style markers. The corpus was POS tagged and function words were extracted. Each text is then represented by two vectors $R_n = \{r_1, r_2, \dots, r_n\}$, one for the normalized frequencies of occurrence of the top 30 function words in the corpus, and another for the normalized frequencies of occurrence of POS-tags. The normalization of the vectors of frequency representing a given text was done

according to the size of the text. Then, for each author, we used 75% of the data generated by texts written by this author to estimate the parameters of the model representing him, and 20% of the data from each author for testing it. The remaining 5% data was merged with 5% of the data (anomalous data) generated by each one of the other authors and was used as a cross-validation set to estimate the probability threshold λ. To get a reasonable estimate of the expected generalization performance, we used resampling with replacement for the two methods. The training and testing process was done 10 times. The overall authorship verification performance is taken as the average performance over these 10 runs. For evaluating the verification performance, we used the standard measures, calculating precision (P), recall (R), and F_1 score.

C.4.3. Baselines

To evaluate the effectiveness of the proposed methods we used one-class SVM as baseline for the unsupervised method and binary SVM classifier as baseline for the weakly supervised method. The one-class SVM was trained and tested on the same data used to train and test the multivariate Gaussian model respectively. The binary SVM classifier was trained on both the data used to train the weakly supervised probabilistic model and the data used to estimate the probability threshold, and it was tested on the same data as the probabilistic model. The overall baselines classification performances are taken as the average performance over the 10 runs as well.

C.5. Results and Discussion

The results of measuring the verification performance for the two different methods in our experimental validation are summarized in what follows. These results show in general the superiority of the proposed methods over the baselines in terms of F_1 on the one hand, and the superiority of the weakly supervised method over the unsupervised method on the other hand. These results also show in general a better performance when using frequent function words than POS-tag for both the proposed method and the baselines.

The preliminary results of measuring the verification performance in our experimental validation for the unsupervised distance-based method against the one-calss SVM are summarized in Table 19. One can notice the clear superiority of the proposed method over the baseline. Our study here indicates that the proposed unsupervised verification method combined with features based on frequent function words can achieve a high verification performance (e.g., $F_1 = 0.83$). As one can expect, increasing the authorship distance threshold ($\propto_2 > \propto_1$) will result in higher recall and lower precision but without significant effect on the F_1 score. By contrast, the one-class SVM performs particularly poorly on this task.

The results of measuring the verification performance for the two different style markers for the weakly supervised against the binary SVM are summarized in Table 20 for function words and in Table 21 for POS tags. These results indicate that the weakly supervised anomaly detection method combined with features based on frequent function words can achieve a high verification performance (e.g., F1 = 0.85).

Table 19. Comparison of average results of the unsupervised authorship verification for the three authors using one-class SVMs and the proposed Unsupervised Anomaly Detection (UAD) method

Method	**P**	**R**	**F_1**
One-class SVMs	0,34	0,50	0,40
UAD ($\propto_1$)	**0,82**	0,85	**0,83**
UAD ($\propto_2$)	0,79	**0,89**	**0,83**

The binary SVM achieved relatively good results but doesn't outperform the probabilistic model; this shows that the authorship verification problem should not be handled as a binary class problem unless a sufficient amount of representative negative data is present to avoid the class imbalance problem. The function words are shown in these results to be more relevant for characterizing the authorial style than POS tags (F1 = 0.85 for function words vs. F1 = 0.77 for POS tags).

Table 20. Results of the weakly supervised authorship verification using frequent function words

Method	**P**	**R**	**F_1**
Binary SVMs	**0,86**	0,75	0,80
Multivariate Gaussian Model	0,82	**0,88**	**0,85**

Table 21. Results of the weakly supervised authorship verification using frequent POS-tags

Method	**P**	**R**	**F_1**
Binary SVMs	**0,81**	0,58	0,67
Multivariate Gaussian Model	0,69	**0,89**	**0,77**

Finally, these results are in line with previous work that claimed that anomaly detection approaches, originating from a supervised classifier (such as one-class SVM), are inappropriate and hardly detect new and unknown anomalies, and that anomaly detection techniques needs to be grounded in the unsupervised learning paradigm (Görnitz et al. 2014). The results suggest also that supervising the anomaly detection process with even a small amount of anomalous data can increase the verification performance.

C.6. A Classic French Literary Mystery: *Le Roman de Violette*

In this last section, we apply our probabilistic approach to settle one of the classic French literary mysteries. *Le Roman de Violette* is a novel published in 1883. The authorship of this novel has still not been determined. Even though the novel was edited under the name of Alexandre Dumas,

some literary critics state that a serious candidate for its authorship is "La Marquise de Mannoury d'Ectot". But this hypothesis cannot be definitely proved, partly because there is only one known book written by that author, which limits the quantity of text available to validate the computational authorship identification methods including our method.

We applied the best performing proposed authorship verification method, which is the weakly supervised one to handle this case. Since there is not enough available text written by "La Marquise de Mannoury d'Ectot" to verify whether she is the writer of *Le Roman de Violette* or not, we set Alexandre Dumas as the author candidate that we want to verify as the writer or not. We trained the probabilistic model based on frequent function words on texts written by Alexandre Dumas. The only known book written by "La Marquise de Mannoury d'Ectot" was used as the representative anomalous text to set the probability threshold. Finally, the verification test was performed on *Le Roman de Violette*. The authorship probability produced by the novel using our proposed method is under the threshold needed to validate the authorship. This result suggests that the novel *Le Roman de Violette* was indeed not written by Alexandre Dumas!

List of Figures

List of Tables

References

Agrawal, R., Imieli'nski, T. & Swami, A., 1993. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*. pp.207–216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*. pp.487–499.

Akmajian, A., Demer, R. A., Farmer, A. K., & Harnish, R. M., 2001. *Linguistics: An introduction to language and communication*. MIT press.

Allport, G.W., 1961. *Pattern and growth in personality*.

Argamon, S., Karlgren, J. & Shanahan, J.G., 2005. *Stylistic analysis of text for information access*, Swedish institute of computer science.

Argamon, S. & Levitan, S., 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F., 2002. An experiment in authorship attribution. In *6th JADT*, pp.29–37.

Beaudouin, V., 2000. Statistique textuelle: une approche empirique du sens à base d'analyse distributionnelle. *Revue Texto*.

Béchet, N., Cellier, P., Charnois, T., & Crémilleux, B., 2012. Discovering linguistic patterns using sequence mining. In *Computational Linguistics and Intelligent Text Processing*, pp.154–165. Springer.

Benzécri, J.-P., 1977. Histoire et préhistoire de l'analyse des données. Partie V: l'analyse des correspondances. *Cahiers de l'analyse des données*, 2(1), pp.9–40.

Berry, D.M., 2011. The computational turn: Thinking about the digital humanities. *Culture Machine*, 12(0), p.2.

Bhattacharyya, P., 2012. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality. *CSI Journal of Computing*, 1(2), pp.1–13.

Biber, D., 2011. Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature*, 1(1), pp.15–23.

Biber, D., 2006. *University language: A corpus-based study of spoken and written registers*, John Benjamins Publishing.

Biber, D. & Conrad, S., 2009. *Register, genre, and style*, Cambridge University Press.

Biber, D., Conrad, S. & Cortes, V., 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3), pp.371–405.

Bordas, É., 2008. *Style: un mot et des discours*. Kimé, p.308

Brachman, R.J. & Anand, T., 1996. The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining*, pp.37–57.

Burrows, J., 2002. "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), pp.267–287.

Burrows, J.F., 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*, Clarendon Pr.

Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A. V, & Oliver, J. L. (2009). Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, *79*(3), 35102.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), p.15.

Charnois, T., Legallois, D. & Larjavaara, M., 2016. *Grammar of Genres and Styles: New approaches*, (To appear).

Chung, C. & Pennebaker, J.W., 2007. The psychological functions of function words. *Social communication*, pp.343–359.

Conesa, G.,1983. *Le dialogue moliéresque*. Presses Universitaires de France

Corbett, E.P.J., 1973. *Classical rhetoric for the modern student*. Oxford University Press

Craig, H., 1999. Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, 14(1), pp.103–113.

Craig, H., 2004. Stylistic analysis and authorship studies. *A companion to digital humanities*, 3, pp.233–334.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P., 1992. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pp.133–140.

De Roeck, A., Sarkar, A. & Garthwaite, P.H., 2004. Defeating the homogeneity assumption. In *Proceedings of 7th International Conference on the Statistical Analysis of Textual Data (JADT)*, pp.282–294.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G., 2003. Authorship attribution with support vector machines. *Applied Intelligence*, *19*(1-2), pp.109–123.

DiMarco, C. & Hirst, G., 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3), pp.451–499.

Eder, M., 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*.

Farris, F. A., 2010. The Gini index and measures of inequality. *American Mathematical Monthly*, *117*(10), pp.851–864.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3), p.37.

Filzmoser, P., 2004. A multivariate outlier detection method. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, pp.18–22.

Fish, S.E., 1979. What is Stylistics and Why Are They Saying Such Terrible Things about It?-Part II. *Boundary 2*, pp.129–146.

Forestier, G. & Bourqui, C., 2010. *Notices de La Jalousie du barbouillé et du Médecin volant* Théâtre co.

Gamon, M., 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*. pp.611.

Ganascia, J.-G., 2001. Extraction of recurrent patterns from stratified ordered trees. In *Machine Learning: ECML 2001*. Springer, pp.167–178.

Ganascia, J.-G., 2002. Extraction of syntatical patterns from parsing trees. In *Internationale Conference on Textual Data Statistical Analysis, 13-15 mars 2002*.

Ganascia, J.-G., 2015. The Logic of the Big Data Turn in Digital Literary Studies. *Frontiers in Digital Humanities*, 2, pp.7.

Garofalakis, M.N., Rastogi, R. & Shim, K., 1999. SPIRIT: Sequential pattern mining with regular expression constraints. In *VLDB*, pp.7–10.

Geng, L. & Hamilton, H.J., 2006. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3), pp.9.

Gildea, D. & Jurafsky, D., 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3), pp.245–288.

Görnitz, N., Kloft, M. M., Rieck, K., & Brefeld, U., 2014. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*.

Grzybek, P., 2014. The Emergence of Stylometry: Prolegomena to the History of Term and Concept. *Text within Text - Culture within Culture*, pp.58–75.

Guiraud, P., 1960. *Problèmes et méthodes de la statistique linguistique*, Presses universitaires de France.

Halliday, M.A.K., 1978. *Language as social semiotic*, London Arnold.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M.-C., 2000. FreeSpan: frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.355–359.

Hawcroft, M., 2007. *Molière: reasoning with fools*, Oxford University Press, USA.

Heller, K., Svore, K., Keromytis, A. D., & Stolfo, S., 2003. One class support vector machines for detecting anomalous windows registry accesses. In *Workshop on Data Mining for Computer Security (DMSEC), Melbourne, FL, November 19, 2003*, pp.2–9.

Holmes, D.I., 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3), pp.111–117.

Holmes, D.I., Robertson, M. & Paez, R., 2001. Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3), pp.315–331.

Hoover, D.L., 2003. Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3), pp.261–286.

Hoover, D.L., 2008. Quantitative analysis and literary studies. *A Companion to Digital Literary Studies*, pp.517–533.

Hovy, E.H., 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2), pp.153–197.

Hu, Y. & Panda, B., 2004. A data mining approach for database intrusion detection. In *Proceedings of the 2004 ACM symposium on Applied computing*, pp.711–716.

Jackendoff, R. & Jackendoff, R.S., 1992. *Semantic structures*, MIT press.

Jockers, M.L., 2013. *Macroanalysis: Digital methods and literary history*, University of Illinois Press.

Johnstone, B., 1996. *The Linguistic Individual: Self-Expression in Language and Linguistics* Oxford University Press, ed.,

Jurafsky, D. & James, H., 2009. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C., 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics*. Vol. 3, pp.255–264.

Kessler, B., Numberg, G. & Schütze, H., 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. pp. 32–38.

Khalaf, Z.A., Alabbas, M. & Tan, T.P., 2011. BASRAH: Arabic Verses Meters Identification System. In *Asian Language Processing (IALP), 2011 International Conference on*. pp.41–44.

Khmelev, D. V & Tweedie, F.J., 2001. Using Markov Chains for Identification of Writer. *Literary and linguistic computing*, 16(3), pp.299–307.

Kirschenbaum, M., 2012. What is digital humanities and what's it doing in English departments? *Debates in the digital humanities*, 3.

Kjellmer, G., 1987. Aspects of English collocations. In *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. pp.133–140.

Koppel, M. & Schler, J., 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*. p. 62.

Koppel, M., Schler, J. & Argamon, S., 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), pp.9–26.

Kress, G., 1988. *Communication and culture: An introduction*, UNSW Press.

Kukushkina, O. V, Polikarpov, A.A. & Khmelev, D.V., 2001. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2), pp.172–184.

Landow, G.P., 1993. *The digital word: Text-based computing in the humanities*, MIT Press.

Lazer, D., Pentland, A. & others, 2009. Computational social science. *Science*, 323, pp.721–723.

Lebart, L., Salem, A. & Berry, L., 1998. *Exploring textual data*, Springer Science & Business Media.

Leech, G.N. & Short, M., 2007. *Style in fiction: A linguistic introduction to English fictional prose*, Pearson Education.

Longrée, D., Luong, X. & Mellet, S., 2008. Les motifs: un outil pour la caractérisation topologique des textes. *S. Heiden et B. Pinceminéds, JADT*, pp.733–744.

Lutoslawski, W., 1898. Principes de stylométrie appliqués a la chonologie des oeuvres de Platon. *Revue des Études Grecques*, 11(41), pp.61-81

Magri-Mourgues, V., 2006. Corpus et stylistique. *Corpus*, (5), pp.5–9.

Mahlberg, M., 2013. *Corpus Stylistics and Dickens' Fiction*, Routledge.

Mahlberg, M., 2005. *English general nouns: A corpus theoretical approach*, John Benjamins Publishing.

Mangiapane, S., 2012. Ponctuation et mise en page dans Madame Bovary: les interventions de Flaubert sur le manuscrit du copiste. *Flaubert. Revue critique et génétique*, (8).

Marcus, M.P., Marcinkiewicz, M.A. & Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), pp.313–330.

Markou, M. & Singh, S., 2003. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12), pp.2481–2497.

Montemurro, M.A., 2001. Beyond the Zipf--Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3), pp.567–578.

Moretti, F., 2005. *Graphs, maps, trees: abstract models for a literary history*, Verso.

Mosteller, F. & Wallace, D.L., 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302), pp.275–309.

Müller, C., 1967. *Étude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille*, French & European Pubns.

O'Connor, B.T., 2014. *Statistical Text Analysis for Social Science*. PhD thesis, Carnegie Mellon University.

O'Neill, M. & Ryan, C., 2001. Grammatical evolution. *IEEE Transactions on Evolutionary Computation*, 5(4), pp.349–358.

Ortuno, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M., 2002. Keyword detection in natural languages and DNA. *EPL (Europhysics Letters)*, *57*(5).

Pei, J. et al., 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *icccn*. pp.215.

Pitler, E. & Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.186–195.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D., 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics? In *Computational Linguistics and Intelligent Text Processing*, pp.166–177. Springer.

Raben, J., 1965. A computer-aided study of literary influence: Milton to Shelley. In *Literary Data Processing Conference Proceedings*, pp.230–274.

Ramsay, S., 2007. Algorithmic criticism. *A Companion to Digital Literary Studies. Blackwell, Oxford*.

Ramsay, S., 2011. *Reading machines: Toward an algorithmic criticism*, University of Illinois Press.

Ramyaa, C.H. & Rasheed, K., 2004. Using machine learning techniques for stylometry. In *Proceedings of International Conference on Machine Learning*.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), pp.846–850.

Rastier, F., 2011. *La mesure et le grain: sémantique de corpus*, Champion; diff. Slatkine.

Rayson, P., 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), pp.519–549.

Renouf, A. & Sinclair, J.M., 2014. 9 Collocational frameworks in English. *English corpus linguistics*, pp.128.

Riffaterre, M., 1971. *Essais de stylistique structurale*, Paris, Flammarion.

Rousseeuw, P.J. & Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), pp.633–639.

Rudman, J., 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4), pp.351–365.

Salem, A., 1986. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, 1(2), pp.5–28.

Sanders, W., 1977. Linguistische Stilistik: Grundżeuge der Stilanalyse sprachlicher Kommunikation.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*. pp.44–49.

Schnapp, J. et al., 2009. The digital humanities manifesto 2.0. *Retrieved September*, 23, 2012.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), pp.1–47.

Semino, E. & Short, M., 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*, Routledge.

Siemens, R. & Schreibman, S., 2013. *A companion to digital literary studies*, John Wiley & Sons.

Simpson, P., 2004. *Stylistics: A resource book for students*, Psychology Press.

Sinclair, J., 1996. Preliminary recommendations on corpus typology. *EAGLES Document.*

Srikant, R. & Agrawal, R., 1996. *Mining sequential patterns: Generalizations and performance improvements*, Springer.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), pp.538–556.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G., 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2), pp.193–214.

Stonebraker, M., Agrawal, R., Dayal, U., Neuhold, E. J., & Reuter, A., 1993. DBMS research at a crossroads: The vienna update. In *VLDB*. Vol. 93, pp.688–692.

Tognini-Bonelli, E., 2001. *Corpus linguistics at work*, John Benjamins Publishing.

Traxler, M.J., 2014. Trends in syntactic parsing: anticipation, Bayesian estimation, and good-enough parsing. *Trends in cognitive sciences*, 18(11), pp.605–611.

Tutin, A., 2009. Showing Phraseology in Context: An Onomasiological Access to Lexico-Grammatical Patterns in Corpora of French Scientific Writings. In *eLexicography in the 21st century : New challenges, new applications*. Louvain-la-Neuve, pp.229.

Vergne, J., 1999. Analyseur linéaire avec dictionnaire partiel, *convention d'utilisation de l'analyseur de Jacques Vergne*.

Viger, P. F., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., & Tseng, V. S., 2014. SPMF: A Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research*, *15*, pp.3389–3393.

Vogel, C. & Lynch, G., 2008. Computational Stylometry: Who's in a Play? In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, pp.169–186.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), pp.236–244.

Weisser, M., 2006. Computational Philology. *(available from http://www.martinweisser.org/publications/comp_phil.pdf).*

Wressnegger, C., Schwenk, G., Arp, D., & Rieck, K., 2013. A close look on n-grams in intrusion detection: anomaly detection vs. classification. In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, pp.67–76.

Xue, N., Ng, H. T., Pradhan, S., Bryant, R. P. C., & Rutherford, A. T., 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of CoNLL* (p. 2).Yan, X., Han, J. & Afshar, R., 2003. CloSpan: Mining closed sequential patterns in large datasets. In *In SDM*, pp.166–177.

Yule, G.U., 1944. *The statistical study of literary vocabulary*, CUP Archive.

Zaki, M.J., 2003. Mining data in bioinformatics. *Handbook of Data Mining*, pp.573–596.

Zaki, M.J., 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2), pp.31–60.

Zhao, Y. & Zobel, J., 2005. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology*. Springer, pp.174–189.

Zheng, R., Li, J., Chen, H., & Huang, Z., 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), pp.378–393.

Zhong, N., Yao, Y.Y.Y. & Ohishima, M., 2003. Peculiarity oriented multidatabase mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4), pp.952–960.